



A Bayesian nonparametric approach for generalized Bradley-Terry models in random environment

Sylvain Le Corff, Matthieu Lerasle, Elodie Vernet

► To cite this version:

Sylvain Le Corff, Matthieu Lerasle, Elodie Vernet. A Bayesian nonparametric approach for generalized Bradley-Terry models in random environment. 2018. hal-01860352

HAL Id: hal-01860352

<https://hal.science/hal-01860352>

Preprint submitted on 24 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Bayesian nonparametric approach for generalized Bradley-Terry models in random environment

Sylvain Le Corff¹, Matthieu Lerasle¹, and Elodie Vernet²

¹Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay.

²Centre de Mathématiques Appliquées, École Polytechnique, Palaiseau.

Abstract

This paper deals with the estimation of the unknown distribution of hidden random variables from the observation of pairwise comparisons between these variables. This problem is inspired by recent developments on Bradley-Terry models in random environment [5] since this framework happens to be relevant to predict for instance the issue of a championship from the observation of a few contests per team [7]. This paper provides three contributions on a Bayesian nonparametric approach to solve this problem. First, we establish contraction rates of the posterior distribution. We also propose a Markov Chain Monte Carlo algorithm to approximately sample from this posterior distribution inspired from a recent Bayesian nonparametric method for hidden Markov models. Finally, the performance of this algorithm are appreciated by comparing predictions on the issue of a championship based on the actual values of the teams and those obtained by sampling from the estimated posterior distribution.

1 Introduction

This paper considers a class of models to deal with pairwise comparison of individuals. The outcome of all the successive comparisons is described by an ordering after all the contestants have met once. A potentially large number of participants are ranked using scores which aggregate the outcomes of their meetings. The objective is to predict this ranking when only a small number of pairwise comparisons has been observed. In generalized Bradley-Terry models, each player $i \geq 1$ is associated with an unknown real parameter V_i characterizing his strength or ability. When contestants i and j face each other, the outcome of their meeting is a random variable taking values in a discrete set \mathcal{X} . For all $x \in \mathcal{X}$, this random variable equals x with probability $K(x, V_i, V_j)$, where the function K is known. The most famous example of such a function K is the Bradley-Terry model [2, 39] in which

$$\mathcal{X} = \{0, 1\}, \quad K(x, V_i, V_j) = \frac{V_i^x V_j^{1-x}}{V_i + V_j}. \quad (1)$$

If $x = 1$, i has beaten j , otherwise, j has beaten i . These models have emerged as key tools to describe pairwise comparisons and have been applied to chess ranking [20], sports [31], animal behaviour [36]. The maximum likelihood estimator (MLE) of the strengths V_i was studied theoretically in [30] when each contestant has faced all the others once. This result has led to interesting

developments in computational statistics to design efficient algorithms to approximate the MLE [4, 19]. The assumption that each contestant has faced all the others is restrictive in some applications and there has been several attempt to weaken it [7, 37]. However, it cannot be completely relaxed as the MLE only exists in the Bradley-Terry model if there exists a path between each pair of players in the oriented graph where an edge is drawn from i to j if i has beaten j [39]. When only a few contests have been observed, it is very likely that the previous condition is not met.

Studying the scores in the American baseball league, [31] introduced a simple model where the parameters V_i are independent and identically distributed (i.i.d.) with common uniform distribution on $[\alpha, 1]$ for some parameter $\alpha > 0$ that has to be estimated. The authors obtained striking results for the prediction of consecutive-game team winning and losing streaks for instance. This motivated in [5] the introduction of the Bradley-Terry model in random environment where the function K is given by (1) and the strengths $(V_i)_{i \geq 1}$ are assumed i.i.d. with common distribution π on $(0, +\infty)$. The authors proved in [5] that in the case of general Bradley-Terry models in random environment, the winner of a championship is the one with maximal strength if and only if the tail of π is sufficiently convex. In other words, interesting predictions regarding the outcome of a tournament can be inferred from π . This is why [7] considered the problem of estimating the distribution of the $(V_i)_{i \geq 1}$ in generalized Bradley-Terry models in random environment, showing that the MLE of this distribution can be defined even when each player has only been involved in 2 contests. They also studied the MLE of π when players meet according to the round-robin scheduling, a widely spread method employed by tournament schedulers. The cornerstone of the work presented in [7] is that risk bounds for the MLE of π follow from the analysis of the likelihood of the following graphical model [7, Lemma 1].

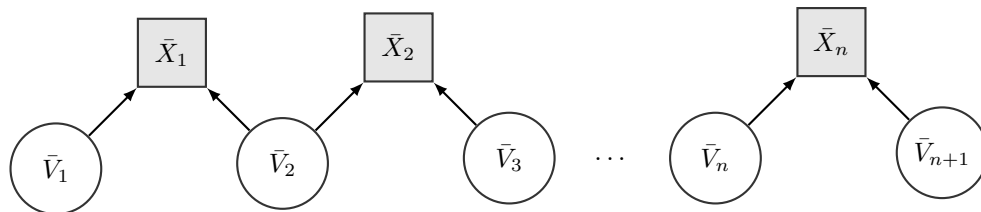


Figure 1: Graphical model of the nonparametric hidden graph.

The details of the link between the graphical model in Figure 1 and (generalized) Bradley-Terry models can be found in [7, Section 2]. Roughly speaking, in Figure 1, each \bar{V}_i , $1 \leq i \leq n+1$, gathers a group $(V_j)_{j \in G_i}$ of contestants in the Bradley-Terry model and each \bar{X}_i , $1 \leq i \leq n$, gathers the outcomes of meetings between contestants j and k when both belong to G_i or when $j \in G_i$ and $k \in G_{i+1}$. In this representation, it is shown in [7] that when meetings are scheduled according to the round-robin algorithm the Markov chain $(\bar{X}_i, \bar{V}_{i+1})_{1 \leq i \leq n}$ is stationary.

The contribution of this paper can be summarized as follows. First, in a Bayesian formulation of the problem considered in [7], Section 3 provides the first nonparametric posterior concentration rates for the estimation of the distribution of the hidden variables $(V_i)_{1 \leq i \leq n+1}$ associated with the graphical model given in Figure 1. Then, Section 4 introduces an algorithm to simulate this posterior distribution. In particular, this algorithm allows to compute an estimator of π and to observe the actual performance of our estimation strategy. These practical aspects are illustrated in a simulation study inspired by data from season 2017-2018 of the French Ligue 1 in Section 4.

In the case of i.i.d. observations, posterior concentration rates have been obtained for a large

class of prior distributions and target densities, see [23, 29] for Gaussian mixtures and [14, 27] for Beta mixtures. Overviews on recent developments of the nonparametric Bayesian theory can be found in [28, 16] and the references therein. Posterior concentration rates were established for partially observed dependent data by [34] following the seminal paper [15]. In [34], the observations arise from a discrete state space hidden Markov model and minimax posterior concentration rates are obtained for the finite-dimensional distributions of the observations. In this paper, posterior concentration rates are obtained for the unknown distribution of latent variables which take values in a continuous state space. Following the approach of [15] (see also [28]) this requires in particular to design exponentially consistent tests by using concentration inequalities for Bradley-Terry models. An important difference with the frequentist analysis performed in [7] is that these inequalities have to be uniform in π to bound the power of these tests. The main results given in Section 3 require therefore a substantial extension of the technical tools used in [7] even if the underlying strategy is similar.

The Bayesian nonparametric procedure introduced in this paper to sample from the posterior distribution of π is inspired by the mixture of Dirichlet processes and the data augmentation scheme presented in [38]. The algorithm proposed by [38] is a Markov Chain Monte Carlo (MCMC) algorithm applied for discrete hidden Markov models with conditional distribution specified as a mixture model in which a base measure is mixed with respect to a Dirichlet process prior (DPP). In Section 4, we propose a mixture of Dirichlet processes formulation where the unknown distribution π is defined as a mixture model where Gaussian distributions are mixed with respect to a DPP. The joint posterior distribution of the parameters of both the DPP and the hidden variables $(V_i)_{1 \leq i \leq n+1}$ is sampled from by adapting the block Gibbs sampling mechanism of [38] to our setting. The main challenge is to sample from the posterior distribution of the states $(V_i)_{1 \leq i \leq n+1}$. Contrary to the discrete case setting considered in [38], exact sampling from this distribution is not possible here and a Sequential Monte Carlo algorithm is used instead. Section 4 presents the performance of the overall procedure when the Forward Filtering Backward Simulation algorithm proposed by [17] (see also [8]) to perform this step is used.

The remaining of the paper is divided as follows. Section 2 displays the formal setting of the paper and Section 3 the main results: posterior concentration rates are established under standard assumptions on the prior distribution. Examples of priors satisfying these assumptions are also given. Section 4 presents the numerical experiments. The MCMC algorithm to sample from the posterior distribution is introduced and an application to a simulated dataset inspired from the results of the French soccer championship is also discussed.

2 Setting

Let n be a positive integer and $(V_i)_{1 \leq i \leq n+1}$ denote i.i.d. random variables taking values in a measurable set $(\mathcal{V}, B(\mathcal{V}))$. For all $1 \leq i \leq n$, the observation X_i takes values in a discrete set \mathcal{X} and conditionally on $(V_i)_{1 \leq i \leq n+1}$ the random variables $(X_i)_{1 \leq i \leq n}$ are independent. The conditional distribution of X_i given $(V_k)_{1 \leq k \leq n+1}$ depends on V_i and V_{i+1} only and is denoted by $x \mapsto K(x, V_i, V_{i+1})$, where $K : \mathcal{X} \times \mathcal{V} \times \mathcal{V} \rightarrow [0, 1]$ is known. Let Π be a set of probability measures on $(\mathcal{V}, B(\mathcal{V}))$. For all $\pi \in \Pi$, note that when the probability distribution of each V_i , $1 \leq i \leq n+1$, is π , then the joint sequence $(X_i, V_{i+1})_{1 \leq i \leq n}$ is a stationary Markov chain with transition kernel

$P : \mathcal{X} \times \mathcal{V} \rightarrow \mathcal{P}(\mathcal{X}) \otimes B(\mathcal{V})$ given, for all $(x, v) \in \mathcal{X} \times \mathcal{V}$ and all $(x', A) \in \mathcal{X} \times B(\mathcal{V})$, by

$$P_\pi(x, v; x', A) = \int \mathbb{1}_A(v') \pi(dv') K(x', v, v'), \quad (2)$$

where $\mathbb{1}_A$ denotes the indicator function of the set A . This Markov chain may be extended to a stationary process $(X_i, V_{i+1})_{i \in \mathbb{Z}}$ with the same transition kernel P_π . In the following, we use the shorthand notation $X = (X_i)_{i \in \mathbb{Z}}$ and $V = (V_i)_{i \in \mathbb{Z}}$. The joint law of this extended joint process is denoted by \mathbb{P}_π and the associated expectation is written \mathbb{E}_π . For all $1 \leq i \leq n$, and all $\pi \in \Pi$,

$$\mathbb{P}_\pi(X_i | V, X_{1:i-1}) = \mathbb{P}_\pi(X_i | V_i, V_{i+1}) = K(X_i, V_i, V_{i+1}),$$

where for any sequence $(a_\ell)_{\ell \in \mathbb{Z}}$, $a_{u:v} = (a_u, \dots, a_v)$ if $u \leq v$ and $a_{u:v} = \emptyset$ if $u > v$. These conditional dependences are represented in the graphical model given in Figure 2. In the Bayesian setting of

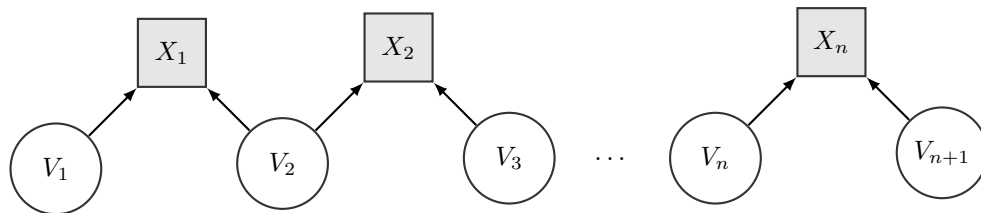


Figure 2: Graphical model of the nonparametric hidden graph.

this paper, the set of distributions Π is equipped with a sigma-algebra \mathcal{A} and a prior distribution μ . The posterior distribution of any $\pi \in \Pi$ given the observations is defined, for all $A \in \mathcal{A}$, by

$$\mu(\pi \in A | X_{1:n}) = \frac{\int \prod_{i=1}^n K(X_i, v_i, v_{i+1}) \mathbb{1}_A(\pi) \pi^{\otimes n+1}(dv_{1:n+1}) \mu(d\pi)}{\int \prod_{i=1}^n K(X_i, v_i, v_{i+1}) \pi^{\otimes n+1}(dv_{1:n+1}) \mu(d\pi)}.$$

In the following, π_\star denotes the true value of the unknown density. This paper focuses on the frequentist properties of the posterior distribution $\mu(\cdot | X_{1:n})$. The aim is to establish rates at which the posterior distribution concentrates around π_\star for a well suited loss function. In the following, let γ be a σ -finite distribution on $(\mathcal{V}, B(\mathcal{V}))$ and S^+ be the set of probability densities with respect to γ . S^+ is equipped with the topology induced by the L_1 -norm and the corresponding Borel σ -field is written $B(S^+)$.

3 Posterior concentration rates

3.1 Main results

Consider the following assumption.

H1 There exists $0 < \nu < 1$ such that for all $x \in \mathcal{X}$ and all $v, w \in \mathcal{V}$, $K(x, v, w) \geq \nu$.

Under H1, the transition kernel of the Markov chain $(V_{i+1}, X_i)_{i \in \mathbb{Z}}$ satisfies, for all $i \in \mathbb{Z}$ and all $A \in B(\mathcal{V})$,

$$P_\pi(X_{i-1}, V_i; X_i, A) = \int \mathbb{1}_A(v_{i+1}) \pi(dv_{i+1}) K(X_i, V_i, v_{i+1}) \geq \nu \pi(A). \quad (3)$$

This uniform lower bound ensures that the joint Markov chain $(V_{i+1}, X_i)_{i \in \mathbb{Z}}$ is uniformly ergodic and that the whole space $\mathcal{V} \times \mathcal{X}$ is a small set. Theorem 3 establishes the posterior concentration rates under \mathbb{P}_{π_*} , it is derived from two important results.

Following [28] and the references therein, Kullback-Leibler type conditions are first required. Kullback-Leibler controls are obtained in Proposition 1. For any $\pi \in \Pi$ and any $x_{1:n} \in \mathcal{X}^n$, let $\ell_n(\pi, x_{1:n})$ be the loglikelihood of the observations $x_{1:n}$,

$$\ell_n(\pi, x_{1:n}) = \log \left(\int \prod_{i=1}^n K(x_i, v_i, v_{i+1}) \pi^{\otimes n+1}(dv_{1:n+1}) \right). \quad (4)$$

Let $\|\cdot\|_{\text{tv}}$ be the total variation distance between probability measures: for all $\pi_1, \pi_2 \in \Pi$,

$$\|\pi_1 - \pi_2\|_{\text{tv}} = \sup_{A \in \mathcal{B}(\mathcal{V})} |\pi_1(A) - \pi_2(A)|.$$

Proposition 1. *Assume that H1 holds. Then, there exists a constant $c_\nu > 0$ such that for all distributions $\pi, \pi' \in \Pi$ and all $n \geq 1$,*

$$\mathbb{E}_\pi [\ell_n(\pi, X_{1:n}) - \ell_n(\pi', X_{1:n})] \leq c_\nu \|\pi - \pi'\|_{\text{tv}}^2 n \quad (5)$$

and

$$\text{Var}_\pi [\ell_n(\pi, X_{1:n}) - \ell_n(\pi', X_{1:n})] \leq c_\nu \|\pi - \pi'\|_{\text{tv}}^2 \log^2(\|\pi - \pi'\|_{\text{tv}}) n. \quad (6)$$

Proof. The proof is postponed to Section A.1. \square

To prove the existence of exponentially consistent tests, required in [28] for example, we use the concentration inequality obtained in Proposition 2. Following [11, Proposition 1] (see also [7]), it is established in Lemma 7 that, under H1, for all $\pi \in \Pi$, there exists a function ℓ_π such that \mathbb{P}_{π_*} -a.s. and in $L^1(\mathbb{P}_{\pi_*})$,

$$\frac{1}{n} \ell_n(\pi, X_{1:n}) \xrightarrow[n \rightarrow \infty]{} \mathbb{L}_{\pi_*}(\pi) = \mathbb{E}_{\pi_*} [\ell_\pi(X)]. \quad (7)$$

The quantity $\ell_\pi(X)$ is the \mathbb{P}_{π_*} -a.s. limit of the sequence $(\log \mathbb{P}_\pi(X_0 | X_{-\ell:-1}))_{\ell \geq 1}$. For all subset $\tilde{\Pi} \subset \Pi$ compact for the topology induced by $\|\cdot\|_{\text{tv}}$, let $\mathbf{N}(\tilde{\Pi}, \|\cdot\|_{\text{tv}}, \varepsilon)$ be the minimal number of balls of radius ε necessary to cover $\tilde{\Pi}$. For all $\pi, \pi' \in \Pi$ and all $n \geq 1$, define

$$\Delta_{\pi, \pi'}^n(X_{1:n}) = \frac{1}{n} \ell_n(\pi', X_{1:n}) - \mathbb{L}_\pi(\pi'). \quad (8)$$

Proposition 2. *Assume that H1 holds. Then, there exists a constant $c_\nu > 0$ such that for all compact subset $\tilde{\Pi} \subset \Pi$ for the topology induced by $\|\cdot\|_{\text{tv}}$, all $\pi \in \tilde{\Pi}$, all $\varepsilon > 0$ and all $t > 0$,*

$$\mathbb{P}_\pi \left(\sup_{\pi', \pi'' \in \tilde{\Pi}} |\Delta_{\pi, \pi'}^n(X_{1:n}) - \Delta_{\pi, \pi''}^n(X_{1:n})| \geq c_\nu \left(\varepsilon + \frac{\sqrt{\log 2\mathbf{N}(\tilde{\Pi}, \|\cdot\|_{\text{tv}}, \varepsilon)} + t}{\sqrt{n}} \right) \right) \leq e^{-t^2}.$$

Proof. The proof is postponed to Section A.2. \square

For all $\varepsilon > 0$, define

$$\mathbf{S}_*(\varepsilon) = \{\pi : \|\pi - \pi_*\|_{\text{tv}} \leq \varepsilon\}. \quad (9)$$

Consider the following assumption.

H2 - (c_1, c_2) There exist a sequence $(\Pi_n)_{n \geq 1}$ of subsets of Π and two sequences $(\varepsilon_n)_{n \geq 1}$ and $(\tilde{\varepsilon}_n)_{n \geq 1}$ such that $(n\tilde{\varepsilon}_n^2)^{-1} \log^2(\tilde{\varepsilon}_n) = o(1)$, $\tilde{\varepsilon}_n \leq \varepsilon_n$ satisfying

$$\begin{aligned} \log N(\Pi_n, \|\cdot\|_{\text{tv}}, \varepsilon_n) &\leq n\varepsilon_n^2, \\ \mu(\mathbf{S}_*(\tilde{\varepsilon}_n)) &\geq e^{-c_1 n \tilde{\varepsilon}_n^2} \quad \text{and} \quad \mu(\Pi_n^c) \leq e^{-c_2 n \tilde{\varepsilon}_n^2}. \end{aligned}$$

As in [7], using that for all $\pi \in \Pi$, $\pi \in \text{Argmax } \mathbf{L}_\pi$, define the risk function on Π as

$$R_{\pi_*} : \pi \mapsto \mathbf{L}_{\pi_*}(\pi_*) - \mathbf{L}_{\pi_*}(\pi)$$

and for all $\varepsilon > 0$ and all constant $\bar{c} > 0$,

$$\mathbf{B}_{\bar{c},*}(\varepsilon) = \{\pi \in \Pi : R_{\pi_*}(\pi) \leq \bar{c}\varepsilon\}.$$

Under assumption H2, posterior concentration rates are derived with respect to the risk function R_{π_*} .

Theorem 3. Assume that H1 and H2- (c_1, c_2) hold for constants $c_1, c_2 > 0$ such that $c_2 > c_\nu + c_1$ where c_ν is defined in Proposition 1. For any positive sequence $(\alpha_n)_{n \geq 1}$ such that $\alpha_n^{-1} = O(e^{c_\alpha n \tilde{\varepsilon}_n^2})$ with $0 < c_\alpha < c_2 - c_\nu - c_1$ and any $\bar{c} > 2(1 + \sqrt{2})(1 + \sqrt{c_2})c_\nu$,

$$\mathbb{P}_{\pi_*}(\mu(\mathbf{B}_{\bar{c},*}^c(\varepsilon_n) | X_{1:n}) > \alpha_n) = o(1).$$

Proof. The proof is postponed to Section A.3. □

Theorem 3 establishes convergence rates of the posterior distribution for the Kullback-Leibler risk R_{π_*} . The Kullback-Leibler divergence between two probability distributions usually scales as the square norm between these probabilities, so that a scaling in ε_n^2 could have been expected in Theorem 3. The reason is that we use the approach of [7] that builds on the analysis of empirical risk minimizers by Vapnik [33] to bound the risk which results in “slow” rates of convergence of the estimators. “Fast” rates (see [32] for a discussion on fast and slow rates in learning) in ε_n^2 could be derived under margin type assumptions [25, 22]. Proving that such margin conditions are satisfied in our framework is an interesting question that goes beyond the scope of this paper.

3.2 Examples of prior distributions

This section presents classical examples of prior distributions satisfying the assumptions of Theorem 3.

Dirichlet process mixtures of Gaussian distributions Dirichlet process mixtures of Gaussian distributions are commonly used to model densities on \mathbb{R} . In [29] and [23], this prior distribution is used to obtain adaptive posterior concentration rates for i.i.d. observations. Let G denote a probability measure on \mathbb{R} , with positive density on \mathbb{R} with respect to the Lebesgue measure and let Π_σ denote a probability measure on $(0, +\infty)$. These measures are assumed to satisfy the following tail assumptions.

H3 $-(\kappa)$ There exist $(b_i)_{i \in \{0,1\}}, x_0 > 0$ such that

$$\begin{aligned} \forall x > x_0, \quad 1 - G([-x, x]) &\leq b_0 e^{-b_1 x^{b_1}}, \\ \forall x < x_0^{-1}, \quad \Pi_\sigma([0, x]) &\leq b_0 e^{-b_1 x^{-b_1}}, \\ \forall x > x_0, \quad \Pi_\sigma([x, +\infty)) &\leq b_0 x^{-b_1}, \\ \forall s > 0, \forall t \in (0, 1), \quad \Pi_\sigma((s, s(1+t))) &\geq b_1 s^{-b_0} t^{b_0} e^{-b_0 s^{-\kappa/2}}. \end{aligned}$$

Let $a > 0$, $\text{DP}(aG)$ be the Dirichlet process with base measure aG , ϕ_σ be the Gaussian density function with mean 0 and variance σ on \mathbb{R} and assume that under the prior distribution

$$\pi = \int \phi_\sigma(\cdot - \mu) dP(\mu), \quad (P, \sigma) \sim \text{DP}(aG) \otimes \Pi_\sigma. \quad (10)$$

Let $\beta > 0$, $\tau \geq 0$, L a non-negative function on \mathbb{R} and let $\mathcal{C}(\beta, L, \tau)$ be the class of locally-Hölder functions with regularity β :

$$\mathcal{C}(\beta, L, \tau) = \left\{ f : \mathbb{R} \rightarrow \mathbb{R} : \forall (x, y) \in \mathbb{R}^2, \left| f^{(\lfloor \beta \rfloor)}(x) - f^{(\lfloor \beta \rfloor)}(y) \right| \leq L(y) e^{\tau|x-y|^2} |x-y|^{\beta-\lfloor \beta \rfloor} \right\},$$

where $\lfloor \beta \rfloor$ denotes the largest integer smaller than β and, for any integer k , $f^{(k)}$ denotes the k -th derivative of f . Consider also the following assumption.

H4 $-(\beta, L)$ There exist positive constants δ, x_m, b_4, b_5 and τ such that

$$\begin{aligned} \forall k \leq \lfloor \beta \rfloor, \quad \int \left(\frac{|\pi_\star^{(k)}(x)|}{\pi_\star(x)} \right)^{(2\beta+\delta)/k} \pi_\star(dx) &< \infty, \\ \int \left(\frac{L(x)}{\pi_\star(x)} \right)^{(2\beta+\delta)/\beta} \pi_\star(dx) &< \infty, \\ \forall |x| > x_m, \quad \pi_\star(x) &\leq b_4 e^{-b_5 |x|^\tau}. \end{aligned}$$

Corollary 4. Let $\pi_\star \in \mathcal{C}(\beta, L, \tau)$. Assume that **H3** $-(\kappa)$ and **H4** $-(\beta, L)$ hold with $\kappa > 0$ and let $d^* = \max(1, \kappa)$. Then, there exist c_1, c_2 such that the prior distribution μ defined in (10) satisfies **H2** $-(c_1, c_2)$ for every positive constants c_1 and c_2 with, for all $n \geq 1$, all $t > t_0 > (d^*(1 + \tau^{-1} + \beta^{-1}) + 1)/(2 + d^*/\beta)$,

$$\varepsilon_n \propto n^{-\beta/(2\beta+d^*)} (\log n)^t, \quad \tilde{\varepsilon}_n \propto n^{-\beta/(2\beta+d^*)} (\log n)^{t_0}.$$

Proof. For all $n \geq 1$, let $\alpha_n = \alpha_0 n^{2/(2\beta+d^*)} (\log n)^{-2t/\beta}$ with $\alpha_0 > 0$ a sufficiently large constant. By [29, Theorem 4],

$$\mu \left(\left\{ \pi : \text{KL}(\pi_\star, \pi) \leq \alpha_n^{-\beta}, V_2(\pi_\star, \pi) \leq \alpha_n^{-\beta} \right\} \right) \geq e^{-c\alpha_n^{1/2} (\log \alpha_n)^{(2+\beta^{-1})t}},$$

where for all $p > 0$,

$$\text{KL}(\pi_\star, \pi) = \int \pi_\star(u) \log \left(\frac{\pi_\star(u)}{\pi(u)} \right) du \quad \text{and} \quad V_p(\pi_\star, \pi) = \int \pi_\star(u) \left| \log \left(\frac{\pi_\star(u)}{\pi(u)} \right) \right|^p du. \quad (11)$$

Then, by Pinsker's inequality, $\|\pi_\star - \pi\|_{\text{TV}} \leq \sqrt{2} \text{KL}(\pi_\star, \pi)^{1/2}$ and therefore

$$\mu \left(S_\star(\sqrt{2}\alpha_n^{-\beta/2}) \right) \geq e^{-c\alpha_n^{1/2} (\log \alpha_n)^{(2+\beta^{-1})t}} \geq e^{-cn\tilde{\varepsilon}_n^2},$$

where S_\star is defined by (9). This concludes the proof using [29, Theorem 5]. \square

Dirichlet process mixtures of Beta distributions Following [27], the posterior concentration rate established in Theorem 3 may be investigated in the case of mixtures of Beta distributions. For all $\alpha > 0$ and all $\varepsilon \in (0, 1)$, the probability density function of the Beta distribution with mean ε and scale parameter α is written:

$$g_{\alpha, \varepsilon} : x \mapsto \frac{x^{a_{\alpha, \varepsilon}-1} (1-x)^{b_{\alpha, \varepsilon}-1}}{B(a_{\alpha, \varepsilon}, b_{\alpha, \varepsilon})},$$

where $a_{\alpha, \varepsilon} = \alpha/(1-\varepsilon)$, $b_{\alpha, \varepsilon} = \alpha/\varepsilon$ and for all $a, b > 0$, $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, with Γ the Gamma function. For any probability distribution P on $(0, 1)$ and any $\alpha > 0$, define the continuous mixture $g_{\alpha, P}$ by

$$g_{\alpha, P} = \int g_{\alpha, \varepsilon} P(d\varepsilon). \quad (12)$$

[27] introduces the following prior distribution on density functions $\pi = g_{\alpha, P}$ on $(0, 1)$ with a random mixture distribution P :

$$(P, \alpha) \sim DP(\nu) \otimes \pi_{\alpha}, \quad (13)$$

where $\pi = g_{\alpha, P}$ and $P = \sum_{j=1}^k p_j \delta_{\varepsilon_j}$. As in [27], assume that:

- η is a probability distribution on $(0, 1)$ such that there exist $\underline{\eta}, \bar{\eta} > 0$ and $T \geq 1$ satisfying, for all $x \in (0, 1)$,

$$0 < \underline{\eta} x^T (1-x)^T \leq \eta(x) \leq \bar{\eta};$$

- the density π_{α} has support $[n^t, +\infty)$ for some $0 < t < 1$ and for all $b_1 > 0$, there exist $c, c_1, c_2, c_3 > 0$ such that for all α_n satisfying $\alpha_n n^{-t} \rightarrow +\infty$,

$$\pi_{\alpha}(c_1 \alpha_n < \alpha < c_2 \alpha_n) \geq c e^{-b_1 \alpha_n^{1/2}} \quad \text{and} \quad \pi_{\alpha}(c_3 \alpha_n < \alpha) \leq c e^{-b_1 \alpha_n^{1/2}}.$$

Let $\beta, L > 0$ and $\mathcal{H}(\beta, L)$ be the class of Hölder functions with regularity β :

$$\mathcal{H}(\beta, L) = \left\{ f : (0, 1) \rightarrow \mathbb{R} : \forall (x, y) \in (0, 1)^2, \left| f^{(\lfloor \beta \rfloor)}(x) - f^{(\lfloor \beta \rfloor)}(y) \right| \leq L |x - y|^{\beta - \lfloor \beta \rfloor} \right\}.$$

Consider also the following assumption.

H5 $-(\beta)$ For all $x \in (0, 1)$ $\pi_{\star}(x) > 0$ and there exist $\beta, L > 0$ and $0 \leq k_0, k_1 < \beta$ such that

$$\pi_{\star} \in \mathcal{H}(\beta, L), \quad \pi_{\star}^{(k_0)}(0) > 0, \quad \pi_{\star}^{(k_1)}(1) < 0.$$

Corollary 5. Assume that **H5** $-(\beta)$ holds with $0 < \beta \leq t - 1/2$. Then, there exist constants c_1, c_2 such that the prior distribution (13) satisfies **H2** $-(c_1, c_2)$ with, for some sufficiently large constant c_0 and all $n \geq 1$,

$$\varepsilon_n = c_0 n^{-\beta/(2\beta+1)} (\log n)^{5\beta/(2\beta+1)}, \quad \tilde{\varepsilon}_n = n^{-\beta/(2\beta+1)} (\log n)^{5\beta/(4\beta+2)}.$$

Proof. In [27], the author analyzed the approximation of the probability density π_{\star} on $(0, 1)$ by a continuous mixture of the form (12). When π_{\star} is Hölder with regularity $\beta > 0$, it is shown in [27, Theorem 3.1] that for all $\alpha > 0$ and all $p > 1$, there exists a probability density π_{α} on $(0, 1)$ such that $\text{KL}(\pi_{\star}, g_{\alpha, \pi_{\alpha}}) \leq c \alpha^{-\beta}$, $V_p(\pi_{\star}, g_{\alpha, \pi_{\alpha}}) \leq c \alpha^{-\beta}$ and $\|\pi_{\star} - g_{\alpha, \pi_{\alpha}}\|_{\infty} \leq \alpha^{-\beta/2}$ where KL and V_p

are defined in (11). If $0 < \beta \leq 2$ we may choose $\pi_\alpha = \pi_\star$ and, if $\beta > 2$, π_α must be different from π to obtain the optimal approximation. Therefore, if π_\star is Hölder with regularity $\beta > 0$, π_\star can be sharply approximated by a continuous mixture of Beta distributions of the form (12). Then, [27] introduced a discrete mixture approximation of g_{α, π_α} of the form g_{α, P_α} where P_α is a probability distribution with finite support. Using this distribution, it is shown in [27, page 171] that

$$\mu \left(\left\{ \pi : \text{KL}(\pi_\star, \pi) \leq \alpha_n^{-\beta}, V_2(\pi_\star, \pi) \leq \alpha_n^{-\beta} \right\} \right) \geq e^{-c\alpha_n^{1/2}(\log \alpha_n)^{5/2}},$$

for some positive constant c . Then, by Pinsker inequality, $\|\pi_\star - \pi\|_{\text{tv}} \leq \sqrt{2} \text{KL}(\pi_\star, \pi)^{1/2}$ and

$$\mu \left(S_\star(\sqrt{2}\alpha_n^{-\beta/2}) \right) \geq e^{-cN_0\alpha_n^{1/2}(\log \alpha_n)^{5/2}} \geq e^{-cn\tilde{\varepsilon}_n^2},$$

where $\tilde{\varepsilon}_n = \alpha_n^{-\beta/2} = \varepsilon_0 n^{-\beta/(2\beta+1)}(\log n)^{5\beta/(4\beta+2)}$ and where S_\star is defined by (9). On the other hand, [27, proof of theorem 2.2, pages 171 to 173] introduces a sequence of subsets $(\Pi_n)_{n \geq 1}$ such that,

$$\mu(\Pi_n^c) \leq C\alpha_n^\beta e^{-\alpha\sqrt{\alpha_n} \log(\alpha_n)^{5/2}} \leq e^{-cn\tilde{\varepsilon}_n^2}$$

and with a log-entropy of Π_n in total variation distance lower than $n\tilde{\varepsilon}_n^2$. \square

4 Numerical experiments

4.1 Block Gibbs sampling algorithm

Following [38], a Markov Chain Monte Carlo (MCMC) method can be introduced to solve the Bayesian nonparametric problem introduced in this paper, i.e. to sample from the posterior distribution given $X_{1:n}$. The unknown distribution π is specified as a mixture model in which some probability density φ_z is mixed with respect to a discrete probability measure P . Defining the discrete probability measure P as in [38], the mixture of Dirichlet processes with base measure αQ , where α is a positive constant and Q a probability distribution, is given by:

$$(\vartheta_j)_{j \geq 1} \underset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha), \quad (14)$$

$$\omega_1 = \vartheta_1 \quad \text{and} \quad \text{for } j \geq 2, \quad \omega_j = \vartheta_j \prod_{i=1}^{j-1} (1 - \vartheta_i), \quad (15)$$

$$(z_j)_{j \geq 1} \underset{\text{i.i.d.}}{\sim} Q, \quad (16)$$

$$(\kappa_i, u_i)_{1 \leq i \leq n+1} \underset{\text{i.i.d.}}{\sim} \sum_{j \geq 1} \mathbb{1}_{u_i < \omega_j} \delta_j(\kappa_i), \quad (17)$$

$$V_i \sim \varphi_{z_{\kappa_i}} \quad \text{for } 1 \leq i \leq n+1, \quad (18)$$

The Dirichlet process prior is defined by the stick-breaking weights $(\vartheta_j)_{j \geq 1}$, the mixture parameters $(z_j)_{j \geq 1}$ the allocation variables $(\kappa_i)_{1 \leq i \leq n+1}$ and some auxiliary variables $(u_i)_{1 \leq i \leq n+1}$. This representation with auxiliary variables is introduced in [35] and used in [38]. For all $1 \leq i \leq n+1$, integrating out the random variable u_i on $(0, 1)$ leads to the usual marginal distribution for κ_i :

$$\kappa_i \underset{\text{i.i.d.}}{\sim} \sum_{j \geq 1} \omega_j \delta_j,$$

and moreover integrating out the random variable κ_i , marginally

$$V_i \sim \sum_{j \geq 1} \omega_j \varphi_{z_j} = \int \varphi_z(\cdot) P(dz),$$

where $P = \sum_{j \geq 1} \omega_j \delta_j$ is distributed from a Dirichlet process with base measure αQ . The auxiliary variables greatly simplifies the expression of the posterior distributions of κ_i as detailed in (ii) below. Sampling from the joint posterior distribution of $(V_{1:n+1}, u_{1:n+1}, \kappa_{1:n+1}, z, \vartheta)$ is then performed by block Gibbs sampling according to the following steps, where $z = (z_j)_{j \geq 1}$ and $\vartheta = (\vartheta_j)_{j \geq 1}$.

- (i) The hidden states $V_{1:n+1}$ are sampled according to their posterior distribution given the random variables $(X_{1:n}, u_{1:n+1}, z, \vartheta)$ with a collapsed step. Note that by integrating out $\kappa_{1:n+1}$, the conditional distribution of $(X_{1:n}, V_{1:n+1})$ given $(u_{1:n+1}, z, \vartheta)$ is given by

$$p_n(V_{1:n+1}, X_{1:n} | u_{1:n+1}, z, \vartheta) \propto \prod_{i=1}^{n+1} \left(\sum_{j, u_i < \omega_j} \varphi_{z_j}(V_i) \right) \prod_{i=1}^n \mathbf{K}(X_i, V_i, V_{i+1}).$$

Therefore, the posterior distribution of $V_{1:n+1}$ given the random variables $(X_{1:n}, u_{1:n+1}, z, \vartheta)$ is the joint smoothing distribution of $V_{1:n+1}$ given $X_{1:n}$ when $(V_i)_{1 \leq i \leq n+1}$ are independent with $V_i \sim \sum_{j, u_i < \omega_j} \varphi_{z_j}$ for all $1 \leq i \leq n+1$. This step cannot be done explicitly as in [38] where the hidden states are discrete. A Sequential Monte Carlo smoother described below is used instead.

- (ii) For all $1 \leq i \leq n+1$, the posterior distribution of κ_i given $(X_{1:n}, V_{1:n+1}, u_{1:n+1}, \vartheta, z)$ is given by

$$\kappa_i | X_{1:n}, V_{1:n+1}, u_{1:n+1}, \vartheta, z \sim \sum_{j, \omega_j > u_i} \varphi_{z_j}(V_i) \delta_j(\kappa_i).$$

This expression motivates the use of the auxiliary variables $u_{1:n+1}$. In the case where $u_{1:n+1}$ are not introduced in the model, κ_i is sampled from its posterior distribution given $(X_{1:n}, \kappa_{1:n+1}, z, \vartheta)$ which is proportional to $\sum_{j \geq 1} \omega_j \varphi_{z_j}(V_i) \delta_j(\kappa_i)$. Due to the infinite sum, this distribution has an intractable normalizing constant so that sampling κ_i without the auxiliary variable u_i is challenging.

- (iii) $(\vartheta, u_{1:n+1})$ are sampled according to their posterior distribution given $(\kappa_{1:n+1}, X_{1:n}, V_{1:n+1}, z)$. As detailed in [38], $(\vartheta_j)_{j \geq 1}$ are updated according to their distribution given $(\kappa_{1:n+1}, X_{1:n}, V_{1:n+1}, z)$:

$$\vartheta_j | \kappa_{1:n+1}, X_{1:n}, V_{1:n+1}, z \underset{\text{ind.}}{\sim} \text{Be} \left(m_j + 1, n + 1 - \sum_{\ell=1}^j m_\ell + \alpha \right),$$

where $m_j = \text{Card}(\{1 \leq i \leq n+1; \kappa_i = j\})$. Then, the random variables $(u_i)_{1 \leq i \leq n+1}$ are conditionally independent given $(\kappa_{1:n+1}, \alpha, X_{1:n}, V_{1:n+1}, z, \vartheta)$ and such that $u_i \sim U(0, \omega_{\kappa_i})$.

- (iv) z are sampled according to their posterior distribution given $(X_{1:n}, \kappa_{1:n}, V_{1:n+1}, u_{1:n+1}, \vartheta)$. This posterior distribution is given, for all $j \geq 1$ by

$$z_j | X_{1:n}, \kappa_{1:n}, V_{1:n+1}, u_{1:n+1}, \vartheta \underset{\text{ind.}}{\sim} Q(z_j) \prod_{i=1, \kappa_i=j}^{n+1} \varphi_{z_j}(V_i).$$

Following [38, Section 4.2.2], the probability density function φ_z is set as a Gaussian probability density with mean μ and variance $1/\lambda$ where $z = (\mu, \lambda)$. The distribution Q is chosen as $Q = \mathcal{N}(0, 1) \otimes \text{Gamma}(1, 1)$, a standard Gaussian distribution for μ which is independent of the precision parameter λ distributed according to a $\text{Gamma}(1, 1)$.

Sequential Monte Carlo smoother for step (i)

A Sequential Monte Carlo smoother is used to sample the hidden states $V_{1:n+1}$ according to their posterior distribution given the random variables $(X_{1:n}, u_{1:n}, z, \vartheta)$. This is the joint smoothing distribution of $V_{1:n+1}$ given $X_{1:n}$ when marginally $(V_i)_{1 \leq i \leq n+1}$ are independent with $V_i \sim \sum_{j, u_i < \omega_j} \varphi_{z_j}$ for all $1 \leq i \leq n+1$.

Particle filtering Particle filtering algorithms are simulation based procedures used to approximate recursively the distributions ϕ_k of V_k given $X_{1:k-1}$ for $k = 1$ to $k = n+1$:

$$\phi_k[h] = \mathbb{E}[h(V_k) | X_{1:k-1}] .$$

The auxiliary particle filter of [24, 26] is the most widely used particle filtering method and is a generalization of other approaches such as the algorithms proposed in [18] and [21]. The auxiliary particle filter produces recursively a set of states $(\xi_k^\ell)_{\ell=1}^M$ associated with importance weights $(\omega_k^\ell)_{\ell=1}^M$ using importance sampling and resampling steps. At $k = 1$, $(\xi_1^\ell)_{\ell=1}^M$ are sampled independently from $\varphi_{z_{\kappa_1}}$ and each particle ξ_1^ℓ is associated with the standard importance sampling weight $\omega_1^\ell = 1/M$. For any bounded and measurable function h defined on \mathbf{X} , the expectation $\phi_1[h]$ is approximated by a self normalized importance sampling estimator:

$$\phi_1^M[h] = M^{-1} \sum_{\ell=1}^M h(\xi_1^\ell) .$$

Then, using $\{(\xi_{k-1}^\ell, \omega_{k-1}^\ell)\}_{\ell=1}^M$, the auxiliary particle filter of [24, 26] samples pairs $\{(I_k^\ell, \xi_k^\ell)\}_{\ell=1}^M$ of superscripts and particles from the instrumental distribution:

$$\pi_k(\ell, x) \propto \omega_{k-1}^\ell \rho_k(\xi_{k-1}^\ell) p_k(\xi_{k-1}^\ell, x)$$

defined on $\{1, \dots, M\} \times \mathcal{V}$, where $(\rho_k(\xi_{k-1}^\ell))_{\ell=1}^M$ are adjustment multiplier weights and p_k is a transition density chosen by the user. In the definition of π_k , the weights $\omega_{k-1}^\ell \rho_k(\xi_{k-1}^\ell)$, $1 \leq \ell \leq M$, are used to duplicate promising particles or discard particles with low importance weights at the previous time step. For $\ell \in \{1, \dots, M\}$, ξ_k^ℓ is associated with the following importance weight:

$$\omega_k^\ell = \frac{\left(\sum_{j, u_k < \omega_j} \varphi_{z_j}(\xi_k^\ell) \right) \mathbf{K}(X_k, \xi_{k-1}^{I_k^\ell}, \xi_k^\ell)}{\rho_k(\xi_{k-1}^\ell) p_k(\xi_{k-1}^{I_k^\ell}, \xi_k^\ell)} . \quad (19)$$

Particle smoothing The Forward Filtering Backward Simulation (FFBSi) algorithm proposed by [17] (see also [6, 8, 13] for its convergence properties and [12] for its computational efficiency) samples backward trajectories from time $n+1$ to time 1 among all the M^{n+2} trajectories which can be built using the particles ξ_k^ℓ , with $1 \leq \ell \leq M$ and $1 \leq k \leq n+1$. Each trajectory is sampled according to the following steps:

- At time $k = n + 1$, sample $J_{n+1} \in \{1 \dots, M\}$ with probabilities proportional to ω_{n+1}^ℓ , for $\ell \in \{1, \dots, M\}$.
- For $k = n$ to $k = 0$, define for all $1 \leq \ell \leq M$,

$$\Lambda_k^M(J_{k+1}, \ell) = \frac{\omega_k^\ell \mathbf{K}(X_k, \xi_k^\ell, \xi_{k+1}^{J_{k+1}})}{\sum_{j=1}^M \omega_k^j \mathbf{K}(X_k, \xi_k^j, \xi_{k+1}^{J_{k+1}})}, \quad 1 \leq \ell \leq M. \quad (20)$$

Then, J_k is sampled in $\{1 \dots, M\}$ with probabilities $\Lambda_k^M(J_{k+1}, \ell)$, $1 \leq \ell \leq M$.

These sampling steps produced a trajectory $(\xi_0^{J_1}, \dots, \xi_{n+1}^{J_{n+1}})$ approximately distributed according to the target joint smoothing distribution. The computational complexity of the FFBSi algorithm grows with M^2 due to the normalizing constant of Λ_k^M , $1 \leq k \leq n$. In the context of this paper, \mathbf{K} is upper bounded by 1 and the acceptance rejection mechanism introduced in [8] to implement the FFBSi algorithm with a complexity which grows linearly with M can be used. In this case, at each time step $1 \leq k \leq n$, the new index J_k is sampled in $\{1, \dots, M\}$ with probabilities proportional to $(\omega_k^\ell)_{\ell=1}^M$ and accepted with probability $\mathbf{K}(X_k, \xi_k^{J_k}, \xi_{k+1}^{J_{k+1}})$.

4.2 Application to soccer results

The algorithm introduced in Section 4.1 is used in this section to estimate the distribution of French soccer teams abilities. Using the outcomes (wins, losses, ties) of the first 30 games during the season 2017-2018, the Expectation Maximization (EM) algorithm for generalized Bradley Terry model with home advantage and ties described in [4] is run to estimate the abilities of the 20 teams with $\mathcal{X} = \{-1, 0, 1\}$ and the distribution

$$\mathbf{K}(1, v_i, v_j) = \frac{\alpha e^{v_i}}{\alpha e^{v_i} + \theta e^{v_j}} \quad \text{and} \quad \mathbf{K}(0, v_i, v_j) = \frac{(\theta^2 - 1) \alpha e^{v_i} e^{v_j}}{(\alpha e^{v_i} + \theta e^{v_j})(\theta \alpha e^{v_i} + e^{v_j})}. \quad (21)$$

The parameters α , characterizing the home advantage (v_i is the home team), and θ , associated with ties, are also estimated with the EM algorithm. The estimates $(\hat{v}_i)_{1 \leq i \leq 20}$, $\hat{\alpha}$ and $\hat{\theta}$ are set as the estimates produced by the algorithm *btemhometies*¹, see Figure 3. The (supposedly) unknown target density π_\star for the Block Gibbs sampling algorithm is then defined as kernel density estimate of the strengths (rescaled around zero) $(\hat{v}_i)_{1 \leq i \leq 20}$ obtained with the function *ksdensity* of Matlab, see Figure 3.

Then $n = 3000$ observations are sampled by first sampling independently $(V_i)_{1 \leq i \leq n+1}$ distributed according to this target density and the observation model (21) with $\hat{\alpha}$ and $\hat{\theta}$. The initial estimate of the Block Gibbs sampling algorithm is set as a mixture of two Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ with μ_1 and μ_2 independent with distribution $\mathcal{N}(0, 1)$ and σ_1^{-2} and σ_2^{-2} independent with distribution Gamma(1, 1). The weight of the first Gaussian distribution is distributed as a Beta(1, α) random variable with $\alpha = 1$.

The Sequential Monte Carlo smoother is set as a standard *Bootstrap filter* i.e. at each time step $1 \leq k \leq n+1$, $\rho_k = 1$ and $p_k(x, x') = \sum_{j, u_k < \omega_j} \varphi_{z_j}(x')$. The number of particles is set to $M = 100$ and 6500 iterations of the Block Gibbs sampling algorithm are performed. The first 5000 samples are discarded as a burn-in period and the estimate of π_\star is set as the average over the last 1500

¹<http://www.stats.ox.ac.uk/~caron/code/bayesbt/index.html>

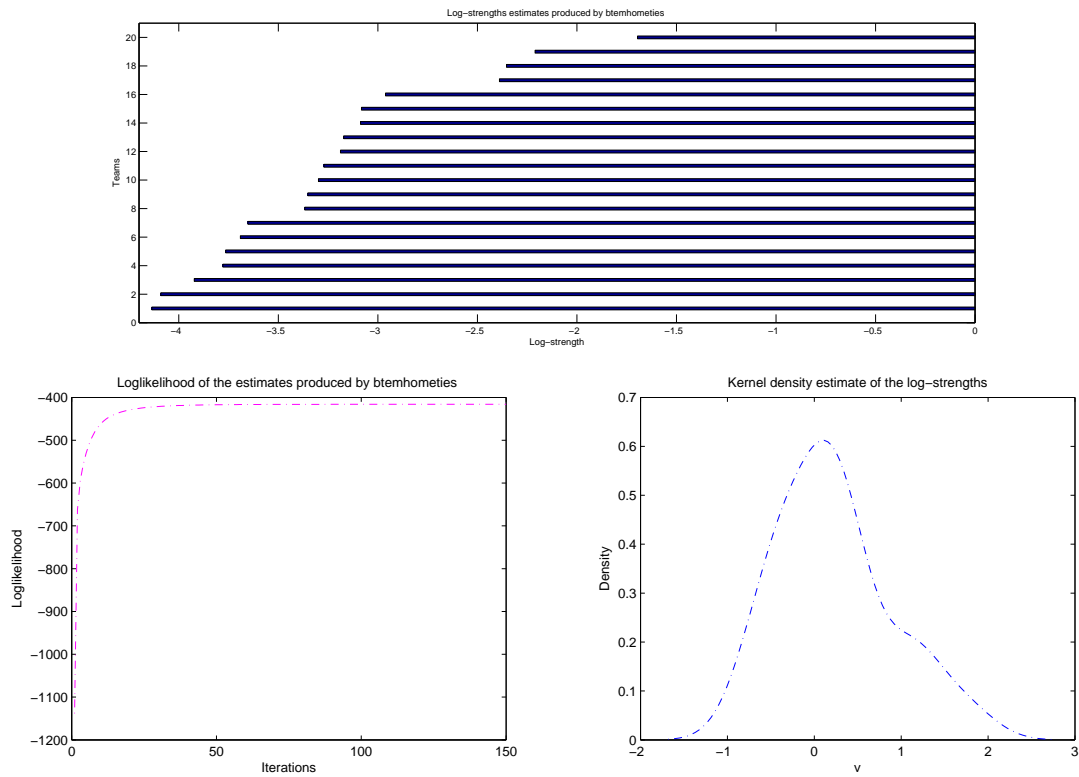


Figure 3: Results of the EM algorithm btmhometies with 2017-2018 Ligue 1 results.

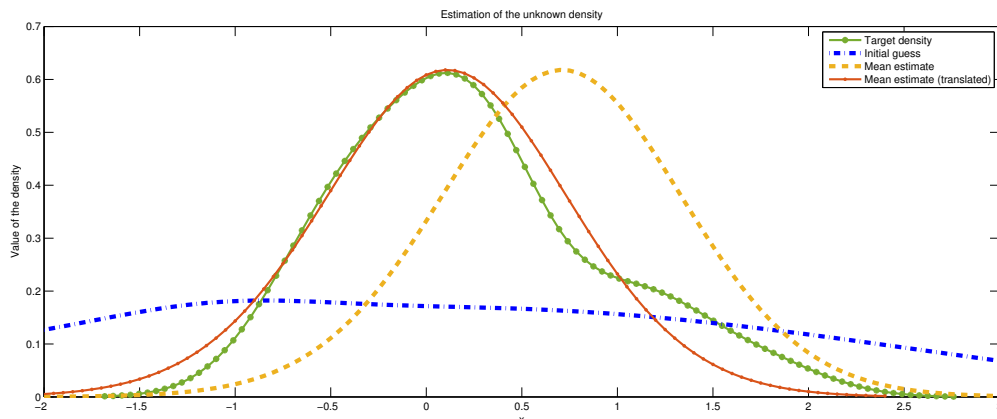


Figure 4: Density estimates from the Block Gibbs sampler.

samples produced by the Block Gibbs Sampler, see Figure 4. By (21), the law of the observations only depends on $(e^{V_i})_{1 \leq i \leq n+1}$, therefore π_* can only be identified up to a translation on \mathbb{R} . Such a translation is displayed in Figure 4 to highlight that the Block Gibbs Sampler recovers mainly the shape of π_* .

In Figure 5 and Figure 6 1000 championships are sampled independently using for each run 20 parameters sampled with the estimated density obtained with the Block Gibbs Sampler to produce boxplots of the scores at the end of the championship. These results are compared to the scores of 1000 championships run with the parameter estimates provided by *btemhometies* in Figure 5 and with parameters sampled from the target density in Figure 6.

A Proof of Theorem 3

A.1 Proof of Proposition 1

Note first that

$$\begin{aligned} \mathbb{E}_\pi [\log \mathbb{P}_\pi(X_{1:n}) - \log \mathbb{P}_{\pi'}(X_{1:n})] &= \sum_{i=1}^n \mathbb{E}_\pi \left[\log \left(\frac{\mathbb{P}_\pi(X_i | X_{1:i-1})}{\mathbb{P}_{\pi'}(X_i | X_{1:i-1})} \right) \right], \\ &= \sum_{i=1}^n \mathbb{E}_\pi \left[\sum_{x_i \in \mathcal{X}} \mathbb{P}_\pi(X_i = x_i | X_{1:i-1}) \log \left(\frac{\mathbb{P}_\pi(X_i = x_i | X_{1:i-1})}{\mathbb{P}_{\pi'}(X_i = x_i | X_{1:i-1})} \right) \right]. \end{aligned}$$

The inner term of the expectation may be upper bounded using that for two probability distributions p and q on \mathcal{X} ,

$$\sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) \leq \sum_{x \in \mathcal{X}} p(x) \frac{p(x) - q(x)}{q(x)} = \sum_{x \in \mathcal{X}} \frac{(p(x) - q(x))^2}{q(x)}.$$

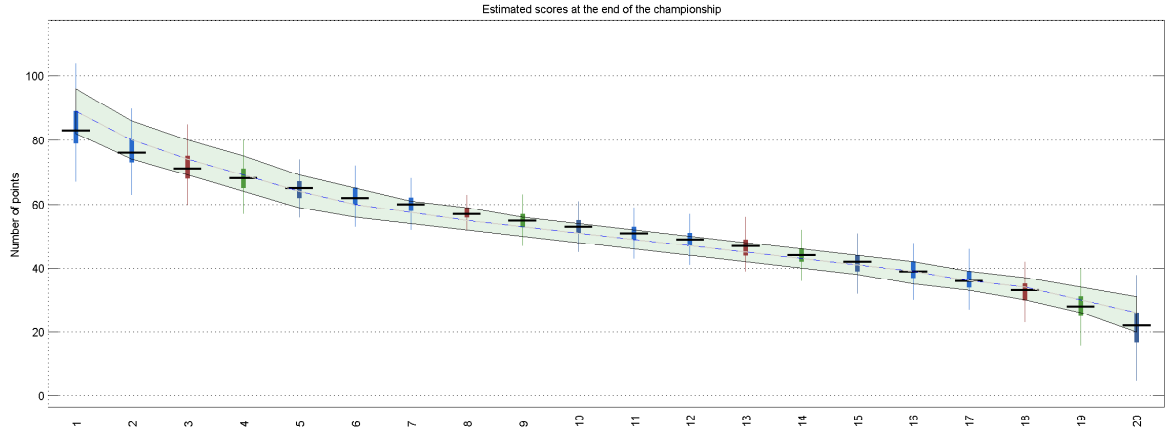


Figure 5: Estimated scores at the end of the championship with *btemhometies* parameter estimates: median (dotted line) and first and last deciles (grey area). Boxplots of the scores obtained with the Block Gibbs Sampler.

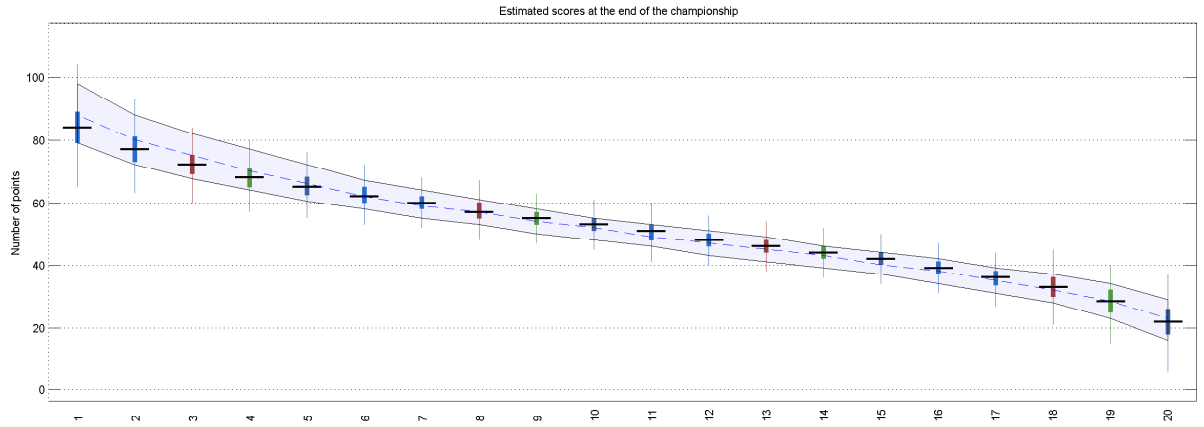


Figure 6: Estimated scores at the end of the championship with parameter sampled with the target distribution: median (dotted line) and first and last deciles (grey area). Boxplots of the scores obtained with the Block Gibbs Sampler.

Then, by Lemma 8,

$$\begin{aligned} \sum_{x_i \in \mathcal{X}} \mathbb{P}_\pi(X_i = x_i | X_{1:i-1}) \log \left(\frac{\mathbb{P}_\pi(X_i = x_i | X_{1:i-1})}{\mathbb{P}_{\pi'}(X_i = x_i | X_{1:i-1})} \right) \\ \leq \sum_{x_i} \frac{(\mathbb{P}_\pi(X_i = x_i | X_{1:i-1}) - \mathbb{P}_{\pi'}(X_i = x_i | X_{1:i-1}))^2}{\mathbb{P}_{\pi'}(X_i = x_i | X_{1:i-1})}, \\ \leq 4|\mathcal{X}|\nu^{-5}(2 + \nu^{-1})^2 \|\pi - \pi'\|_{\text{tv}}^2, \end{aligned}$$

which concludes the proof of (5). By [34, Equation (32)],

$$\begin{aligned} \text{Var}_\pi [\log \mathbb{P}_\pi(X_{1:n}) - \log \mathbb{P}_{\pi'}(X_{1:n})] \\ \leq 4 \sum_{1 \leq i \leq n} \mathbb{E}_\pi [Z_i^2] + 4 \sum_{1 \leq i < j \leq n} \mathbb{E}_\pi [Z_j^2]^{1/2} \mathbb{E}_\pi \left[(\mathbb{E}_\pi [Z_j | X_{1:i-1}] - \mathbb{E}_\pi [Z_j])^2 \right]^{1/2}, \end{aligned} \quad (22)$$

where, for all $i \geq 1$, $Z_i = \log \mathbb{P}_\pi(X_i | X_{1:i-1}) - \log \mathbb{P}_{\pi'}(X_i | X_{1:i-1})$. Following [9], (22) can be upper bounded using the extended Markov chain defined, for all $1 \leq i \leq n$, by

$$R_i = (V_{i+1}, X_i, \eta_i^\pi, \eta_i^{\pi_*}), \quad (23)$$

where, for all $i \geq 1$, η_i^π is the predictive filter at time i when the law of the hidden states is π : $\eta_1^\pi = \pi$ and for $i \geq 2$,

$$\eta_i^\pi : A \mapsto \mathbb{P}_\pi(V_i \in A | X_{1:i-1}).$$

Define the function h , for all $r = (v, x, \eta, \eta_*) \in \mathcal{X} \times \mathcal{V} \times \mathbf{S}^+ \times \mathbf{S}^+$, by:

$$h(r) = \log \left(\frac{\int \pi_*(dv_{p+1}) \eta_*(dv_p) \mathbf{K}(v_p, v_{p+1}, x)}{\int \pi(dv_{p+1}) \eta(dv_p) \mathbf{K}(v_p, v_{p+1}, x)} \right). \quad (24)$$

Then, for all $i \geq 1$,

$$h(R_i) = \log \mathbb{P}_{\pi_*}(X_i | X_{1:i-1}) - \log \mathbb{P}_\pi(X_i | X_{1:i-1}) = Z_i$$

and for all $1 \leq i < j \leq n$,

$$|\mathbb{E}_\pi [Z_j | X_{1:i-1}] - \mathbb{E}_\pi [Z_j]| = |\mathbb{E}_\pi [h(R_j) | X_{1:i-1}] - \mathbb{E}_\pi [h(R_j)]|.$$

Let Φ be defined for any $\pi \in \Pi$, $x \in \mathcal{X}$ and probability distribution η on \mathcal{V} by:

$$\Phi(x, \eta; \pi) : A \mapsto \frac{\int \mathbb{1}_A(z) \pi(dz) \eta(dz') \mathbf{K}(z', z, x)}{\int \pi(dz) \eta(dz') \mathbf{K}(z', z, x)}.$$

For all $i \geq 2$, the predictive filter at time i may be expressed as:

$$\eta_i^\pi = \Phi(X_{i-1}, \eta_{i-1}^\pi; \pi). \quad (25)$$

The transition kernel of the extended chain is given, for all $(x, v, \eta, \eta_*) \in \mathcal{V} \times \mathcal{X} \times \mathbf{S}^+ \times \mathbf{S}^+$ and all $B(\mathcal{V}) \times \mathcal{P}(\mathcal{X}) \times B(\mathbf{S}^+) \times B(\mathbf{S}^+)$, by

$$\begin{aligned} Q_\pi(v, x, \eta, \eta_*, A_\mathcal{V}, A_\mathcal{X}, B_1, B_2) \\ = \sum_{x' \in \mathcal{X}} \int \pi(dv') \mathbf{K}(v, v', x') \mathbb{1}_{A_\mathcal{X}}(x') \mathbb{1}_{A_\mathcal{V}}(v') \mathbb{1}_{B_1}(\Phi(x, \eta; \pi)) \mathbb{1}_{B_2}(\Phi(x, \eta_*; \pi_*)). \end{aligned} \quad (26)$$

For all $p \geq 1$, the p -th iterate of Q_π is denoted by Q_π^p . For all $1 \leq i < j \leq n$, using Lemma 10,

$$\begin{aligned} |\mathbb{E}_\pi [Z_j | X_{1:i-1}] - \mathbb{E}_\pi [Z_j]| &\leq \int \|Q_\pi^{j-i} \mathbf{h}(r) - Q_\pi^{j-i} \mathbf{h}(\tilde{r})\| p_\pi(dr_i | X_{1:i-1}) p_\pi(d\tilde{r}_i), \\ &\leq 6\nu^{-1}(1-\nu)^{j-i-2}. \end{aligned} \quad (27)$$

On the other hand, by Lemma 8, for all $1 \leq j \leq n$,

$$\mathbb{E}_\pi [Z_j^2] \leq 4\nu^{-6}(2+\nu^{-1})^2 \|\pi - \pi'\|_{\text{tv}}^2, \quad (28)$$

so that for all $0 < \beta < 1$, using (22), (27) and (28),

$$\begin{aligned} \text{Var}_\pi [\log \mathbb{P}_\pi(X_{1:n}) - \log \mathbb{P}_{\pi'}(X_{1:n})] &\leq c_\nu \sum_{1 \leq i \leq n} \mathbb{E}_\pi [Z_i^2] + c_\nu \sum_{1 \leq i < j \leq n} \sqrt{\mathbb{E}_\pi [Z_j^2]} \sqrt{|\mathbb{E}_\pi [Z_j | X_{1:i-1}] - \mathbb{E}_\pi [Z_j]|^\beta \mathbb{E}_\pi [Z_j^2]^{1-\beta/2}}, \\ &\leq c_\nu \beta^{-2} \|\pi - \pi'\|_{\text{tv}}^{2-\beta/2} n, \end{aligned}$$

where c_ν is a constant which depends on ν only. The proof is completed by noting that the best upper bound is obtained for $\beta = -4/\log(\|\pi - \pi'\|_{\text{tv}})$.

A.2 Proof of Proposition 2

For all $\pi, \pi', \pi'' \in \tilde{\Pi}$, by (8),

$$\begin{aligned} \Delta_{\pi, \pi'}^n(X_{1:n}) - \Delta_{\pi, \pi''}^n(X_{1:n}) &= \frac{1}{n} \ell_n(\pi', X_{1:n}) - \frac{1}{n} \ell_n(\pi'', X_{1:n}) - \mathbb{L}_\pi(\pi') + \mathbb{L}_\pi(\pi''), \\ &= \frac{1}{n} \ell_n(\pi', X_{1:n}) - \mathbb{E}_\pi \left[\frac{1}{n} \ell_n(\pi', X_{1:n}) \right] + \mathbb{E}_\pi \left[\frac{1}{n} \ell_n(\pi', X_{1:n}) \right] - \mathbb{L}_\pi(\pi') \\ &\quad - \frac{1}{n} \ell_n(\pi'', X_{1:n}) + \mathbb{E}_\pi \left[\frac{1}{n} \ell_n(\pi'', X_{1:n}) \right] - \mathbb{E}_\pi \left[\frac{1}{n} \ell_n(\pi'', X_{1:n}) \right] + \mathbb{L}_\pi(\pi''). \end{aligned}$$

Note first that by Lemma 7, for all $\pi \in \tilde{\Pi}$,

$$\sup_{\pi' \in \tilde{\Pi}} \left| \mathbb{E}_\pi \left[\frac{1}{n} \ell_n(\pi', X_{1:n}) \right] - \mathbb{L}_\pi(\pi') \right| \leq \frac{1}{n\nu^2}.$$

Therefore, for all $\pi \in \tilde{\Pi}$,

$$\sup_{\pi', \pi'' \in \tilde{\Pi}} |\Delta_{\pi, \pi'}^n(X_{1:n}) - \Delta_{\pi, \pi''}^n(X_{1:n})| \leq \frac{2}{n\nu^2} + 2G_\pi^n(X_{1:n}), \quad (29)$$

where

$$G_\pi^n(X_{1:n}) = \sup_{\pi' \in \tilde{\Pi}} |\Delta_{\pi, \pi'}^n(X_{1:n})|, \quad \Delta_{\pi, \pi'}^n(X_{1:n}) = n^{-1} \ell_n(\pi', X_{1:n}) - n^{-1} \mathbb{E}_\pi [\ell_n(\pi', X_{1:n})].$$

G_π^n is a function of the Markov chain $(X_i, V_{i+1})_{1 \leq i \leq n}$ whose transition kernel P_π is uniformly lower bounded by (3). By Corollary 12, it is enough to obtain a bounded difference inequality to establish a concentration inequality for G_π^n . For all $1 \leq k \leq n$ and all $x_{1:n} \in \mathcal{X}^n$ define $\tilde{x}^{(k)} \in \mathcal{X}^n$ such that

$$\tilde{x}_k^{(k)} \neq x_k \quad \text{and for all } j \neq k, \tilde{x}_j^{(k)} = x_j. \quad (30)$$

Note that, for any $\pi' \in \tilde{\Pi}$,

$$\begin{aligned} \left| Z_{\pi, \pi'}^n(x_{1:n}) - Z_{\pi, \pi'}^n(\tilde{x}_{1:n}^{(k)}) \right| &\leq n^{-1} \left| \ell_n(\pi', x_{1:n}) - \ell_n(\pi', \tilde{x}_{1:n}^{(k)}) \right|, \\ &\leq n^{-1} \sum_{i=1}^n \left| \log \mathbb{P}_{\pi'}(x_i | x_{1:i-1}) - \log \mathbb{P}_{\pi'}(\tilde{x}_i^{(k)} | \tilde{x}_{1:i-1}^{(k)}) \right|, \\ &\leq n^{-1} \sum_{i=k}^n \left| \log \mathbb{P}_{\pi'}(x_i | x_{1:i-1}) - \log \mathbb{P}_{\pi'}(\tilde{x}_i^{(k)} | \tilde{x}_{1:i-1}^{(k)}) \right|, \end{aligned}$$

as for $1 \leq i \leq k-1$, $x_i = \tilde{x}_i^{(k)}$. In the case $i = k$,

$$|\mathbb{P}_{\pi'}(x_i | x_{1:i-1}) - \mathbb{P}_{\pi'}(\tilde{x}_i^{(k)} | \tilde{x}_{1:i-1}^{(k)})| = |\mathbb{P}_{\pi'}(x_k | x_{1:k-1}) - \mathbb{P}_{\pi'}(\tilde{x}_k^{(k)} | x_{1:k-1})| \leq 1,$$

so that $|\log \mathbb{P}_{\pi'}(x_i | x_{1:i-1}) - \log \mathbb{P}_{\pi'}(\tilde{x}_i^{(k)} | \tilde{x}_{1:i-1}^{(k)})| \leq \nu^{-1}$. For all $k+1 \leq i \leq n$, by Lemma 6,

$$\mathbb{P}_{\pi'}(\tilde{x}_i^{(k)} | \tilde{x}_{1:i-1}^{(k)}) = \int \mathbb{P}_{\pi'}(dv_{k+1} | \tilde{x}_{1:i-1}^{(k)}) \prod_{j=k+1}^{i-1} K_{\pi', j, i-1}^{V|X}(v_j, dv_{j+1}) \pi'(dv_{i+1}) \mathbf{K}(v_i, v_{i+1}, x_i)$$

and

$$\mathbb{P}_{\pi'}(x_i | x_{1:i-1}) = \int \mathbb{P}_{\pi'}(dv_{k+1} | x_{1:i-1}) \prod_{j=k+1}^{i-1} K_{\pi', j, i-1}^{V|X}(v_j, dv_{j+1}) \pi'(dv_{i+1}) \mathbf{K}(v_i, v_{i+1}, x_i).$$

Therefore, $|\mathbb{P}_{\pi'}(\tilde{x}_i^{(k)} | \tilde{x}_{1:i-1}^{(k)}) - \mathbb{P}_{\pi'}(x_i | x_{1:i-1})| \leq (1 - \nu)^{i-k-1}$ and

$$|\log \mathbb{P}_{\pi'}(\tilde{x}_i^{(k)} | \tilde{x}_{1:i-1}^{(k)}) - \log \mathbb{P}_{\pi'}(x_i | x_{1:i-1})| \leq \nu^{-1} (1 - \nu)^{i-k-1} \quad (31)$$

which yields, for any $\pi' \in \tilde{\Pi}$,

$$\left| Z_{\pi, \pi'}^n(x_{1:n}) - Z_{\pi, \pi'}^n(\tilde{x}_{1:n}^{(k)}) \right| \leq n^{-1} \left(\nu^{-1} + \sum_{i=k+1}^n \nu^{-1} (1 - \nu)^{i-k-1} \right) \leq \frac{2}{n\nu^2}.$$

By Corollary 12 applied with $\gamma_1 = \dots = \gamma_n = 2/(n\nu^2)$, for all $\pi, \pi' \in \tilde{\Pi}$ and all $t > 0$,

$$\mathbb{P}_{\pi} \left(\left| Z_{\pi, \pi'}^n(X_{1:n}) \right| \geq \frac{2\sqrt{5}t}{\nu^3\sqrt{n}} \right) \leq 2e^{-t^2}. \quad (32)$$

By Lemma 8,

$$|Z_{\pi, \pi_1}(x_{1:n}) - Z_{\pi, \pi_2}(x_{1:n})| \leq 4\nu^{-3} (2 + \nu^{-1}) \|\pi_1 - \pi_2\|_{\text{tv}}. \quad (33)$$

Let $\mathbf{R}(\varepsilon)$ denote an ε -net of $\tilde{\Pi}$. By (33),

$$G_{\pi}^n(X_{1:n}) \leq \max_{\pi' \in \mathbf{R}(\varepsilon)} |Z_{\pi, \pi'}(x_{1:n})| + 4\nu^{-3} (2 + \nu^{-1}) \varepsilon.$$

Applying a union bound to (32) yields

$$\mathbb{P}_{\pi} \left(G_{\pi}^n(X_{1:n}) \geq \frac{2\sqrt{5}t}{\nu^3\sqrt{n}} + 4\nu^{-3} (2 + \nu^{-1}) \varepsilon \right) \leq 2N(\tilde{\Pi}, \|\cdot\|_{\text{tv}}, \varepsilon) e^{-t^2}.$$

A.3 Proof of Theorem 3

For $n \geq 1$,

$$\mu(\mathbb{B}_{\bar{c},\star}^c(\varepsilon_n)|X_{1:n}) = \frac{\int_{\mathbb{B}_{\bar{c},\star}^c(\varepsilon_n)} \exp\{\ell_n(\pi, X_{1:n}) - \ell_n(\pi_\star, X_{1:n})\} \mu(d\pi)}{\int_{\Pi} \exp\{\ell_n(\pi, X_{1:n}) - \ell_n(\pi_\star, X_{1:n})\} \mu(d\pi)} = \frac{N_n}{D_n}.$$

Following [28, Section 3.3], consider, for all $u > 0$ and all function $g : \mathbb{R}_+^* \rightarrow \mathbb{R}_+^*$, the decomposition

$$\begin{aligned} \mathbb{P}_{\pi_\star}(\mu(\mathbb{B}_{\bar{c},\star}^c(\varepsilon_n)|X_{1:n}) > \alpha_n) &\leq \mathbb{E}_{\pi_\star} \left[\Phi_n^{\tau_{n,c}(u)}(X_{1:n}, \pi_\star) \right] + \mathbb{P}_{\pi_\star}(D_n < g(u)) \\ &\quad + \mathbb{P}_{\pi_\star} \left(\left\{ \Phi_n^{\tau_{n,c}(u)}(X_{1:n}, \pi_\star) = 0 \right\} \cap \{D_n \geq g(u)\} \cap \left\{ \frac{N_n}{D_n} > \alpha_n \right\} \right), \end{aligned}$$

where for all $n \geq 1$ and all $t > 0$,

$$\Phi_n^t(X_{1:n}, \pi_\star) = \mathbb{1}_{\{\exists \pi \in \Pi_n : \ell_n(\pi, X_{1:n})/n - \ell_n(\pi_\star, X_{1:n})/n + \mathbb{L}_{\pi_\star}(\pi_\star) - \mathbb{L}_{\pi_\star}(\pi) > t\}} \quad (34)$$

and for all $u > 0$, $c > 0$ and $n \geq 1$

$$\tau_{n,c}(u) = c \left(n^{-1/2}u + \varepsilon_n \right).$$

By Markov inequality, for all $u > 0$,

$$\begin{aligned} \mathbb{P}_{\pi_\star} \left(\left\{ \Phi_n^{\tau_{n,c}(u)}(X_{1:n}, \pi_\star) = 0 \right\} \cap \{D_n \geq g(u)\} \cap \left\{ \frac{N_n}{D_n} > \alpha_n \right\} \right) \\ &\leq \mathbb{P}_{\pi_\star} \left(\left\{ \Phi_n^{\tau_{n,c}(u)}(X_{1:n}, \pi_\star) = 0 \right\} \cap \{N_n > \alpha_n g(u)\} \right), \\ &\leq \mathbb{P}_{\pi_\star} \left(\left\{ \int_{\mathbb{B}_{\bar{c},\star}^c(\varepsilon_n)} \left(1 - \Phi_n^{\tau_{n,c}(u)}(X_{1:n}, \pi_\star) \right) \frac{\exp\{\ell_n(\pi, X_{1:n})\}}{\exp\{\ell_n(\pi_\star, X_{1:n})\}} \mu(d\pi) > \alpha_n g(u) \right\} \right), \\ &\leq \frac{1}{\alpha_n g(u)} \mathbb{E}_{\pi_\star} \left[\int_{\mathbb{B}_{\bar{c},\star}^c(\varepsilon_n)} \left(1 - \Phi_n^{\tau_{n,c}(u)}(X_{1:n}, \pi_\star) \right) \frac{\exp\{\ell_n(\pi, X_{1:n})\}}{\exp\{\ell_n(\pi_\star, X_{1:n})\}} \mu(d\pi) \right], \\ &\leq \frac{1}{\alpha_n g(u)} \int_{\mathbb{B}_{\bar{c},\star}^c(\varepsilon_n)} \mathbb{E}_{\pi} \left[1 - \Phi_n^{\tau_{n,c}(u)}(X_{1:n}, \pi_\star) \right] \mu(d\pi). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{P}_{\pi_\star}(\mu(\mathbb{B}_{\bar{c},\star}^c(\varepsilon_n)|X_{1:n}) > \alpha_n) &\leq \mathbb{E}_{\pi_\star} \left[\Phi_n^{\tau_{n,c}(u)}(X_{1:n}, \pi_\star) \right] + \mathbb{P}_{\pi_\star}(D_n < g(u)) \\ &\quad + \frac{1}{\alpha_n g(u)} \int_{\mathbb{B}_{\bar{c},\star}^c(\varepsilon_n) \cap \Pi_n} \mathbb{E}_{\pi} \left[1 - \Phi_n^{\tau_{n,c}(u)}(X_{1:n}, \pi_\star) \right] \mu(d\pi) + \frac{\mu(\Pi_n^c)}{\alpha_n g(u)}. \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{E}_{\pi_\star} \left[\Phi_n^{\tau_{n,c}(u)}(X_{1:n}, \pi_\star) \right] \\ &= \mathbb{P}_{\pi_\star}(\exists \pi \in \Pi_n : (\ell_n(\pi, X_{1:n}) - \ell_n(\pi_\star, X_{1:n}))/n - (\mathbb{L}_{\pi_\star}(\pi) - \mathbb{L}_{\pi_\star}(\pi_\star)) > \tau_{n,c}(u)), \\ &\leq \mathbb{P}_{\pi_\star}(\exists \pi \in \Pi_n : |\Delta_{\pi_\star, \pi}^n(X_{1:n}) - \Delta_{\pi_\star, \pi_\star}^n(X_{1:n})| > \tau_{n,c}(u)), \end{aligned}$$

where $\Delta_{\pi_*, \pi}^n(X_{1:n})$ and $\Delta_{\pi_*, \pi_*}^n(X_{1:n})$ are defined by (8). Then, by Proposition 2 with $\tilde{\Pi} = \Pi_n \cup \{\pi_*\}$, choosing $c = (1 + \sqrt{2})c_\nu$ and $u = u_n = (\bar{c}/(2c) - 1)\sqrt{n}\tilde{\varepsilon}_n$, for any sufficiently large n ,

$$\mathbb{E}_{\pi_*} \left[\Phi_n^{\tau_{n,c}(u_n)}(X_{1:n}, \pi_*) \right] \leq e^{-u_n^2}.$$

This yields

$$\begin{aligned} \mathbb{P}_{\pi_*} \left(\mu(B_{\bar{c},*}^c(\varepsilon_n) | X_{1:n}) > \alpha_n \right) &\leq e^{-u_n^2} + \mathbb{P}_{\pi_*} (D_n < g(u_n)) \\ &\quad + \frac{1}{\alpha_n g(u_n)} \int_{B_{\bar{c},*}^c(\varepsilon_n) \cap \Pi_n} \mathbb{E}_\pi \left[1 - \Phi_n^{\tau_{n,c}(u_n)}(X_{1:n}, \pi_*) \right] \mu(d\pi) + \frac{\mu(\Pi_n^c)}{\alpha_n g(u_n)}. \end{aligned}$$

Then, using that for all $\pi \in \Pi_n$, $L_\pi(\pi) - L_\pi(\pi_*) \geq 0$, and for all $\pi \in B_{\bar{c},*}^c(\varepsilon_n)$, $L_{\pi_*}(\pi_*) - L_{\pi_*}(\pi) > \bar{c}\varepsilon_n$,

$$\begin{aligned} \mathbb{E}_\pi \left[1 - \Phi_n^{\tau_{n,c}(u_n)}(X_{1:n}, \pi_*) \right] &= \mathbb{P}_\pi (\forall \pi' \in \Pi_n : \ell_n(\pi', X_{1:n})/n - \ell_n(\pi_*, X_{1:n})/n - (L_{\pi_*}(\pi') - L_{\pi_*}(\pi_*)) \leq \tau_{n,c}(u_n)) , \\ &\leq \mathbb{P}_\pi (\ell_n(\pi, X_{1:n})/n - \ell_n(\pi_*, X_{1:n})/n \leq \tau_{n,c}(u_n) - \bar{c}\varepsilon_n) , \\ &\leq \mathbb{P}_\pi ((\ell_n(\pi, X_{1:n}) - \ell_n(\pi_*, X_{1:n}))/n - (L_\pi(\pi) - L_\pi(\pi_*)) \leq \tau_{n,c}(u_n) - \bar{c}\varepsilon_n) . \end{aligned}$$

Since $\tau_{n,c}(u_n) \leq \bar{c}\varepsilon_n/2$,

$$\begin{aligned} \mathbb{E}_\pi \left[1 - \Phi_n^{\tau_{n,c}(u_n)}(X_{1:n}, \pi_*) \right] &\leq \mathbb{P}_\pi (|\Delta_{\pi,\pi}^n(X_{1:n}) - \Delta_{\pi,\pi_*}^n(X_{1:n})| \geq \bar{c}\varepsilon_n/2) , \\ &\leq \mathbb{P}_\pi (|\Delta_{\pi,\pi}^n(X_{1:n}) - \Delta_{\pi,\pi_*}^n(X_{1:n})| \geq \tau_{n,c}(u_n)) , \\ &\leq e^{-u_n^2}, \end{aligned}$$

where the last inequality follows from Proposition 2. Then,

$$\mathbb{P}_{\pi_*} \left(\mu(B_{\bar{c},*}^c(\varepsilon_n) | X_{1:n}) > \alpha_n \right) \leq e^{-u_n^2} + \mathbb{P}_{\pi_*} (D_n < g(u_n)) + \frac{e^{-u_n^2}}{\alpha_n g(u_n)} + \frac{\mu(\Pi_n^c)}{\alpha_n g(u_n)}.$$

Let $c_3 > c_\nu$ where c_ν is given in Proposition 1. Since

$$D_n \geq \int_{S_*(\tilde{\varepsilon}_n)} \exp \{ \ell_n(\pi, X_{1:n}) - \ell_n(\pi_*, X_{1:n}) \} \mathbb{1}_{\{ \ell_n(\pi, X_{1:n}) - \ell_n(\pi_*, X_{1:n}) \geq -c_3 n \tilde{\varepsilon}_n^2 \}} \mu(d\pi),$$

then

$$\begin{aligned} \mathbb{P}_{\pi_*} (D_n < g(u_n)) &\leq \mathbb{P}_{\pi_*} \left(\mu(S_*(\tilde{\varepsilon}_n) \cap \{ \pi : \ell_n(\pi, X_{1:n}) - \ell_n(\pi_*, X_{1:n}) \geq -c_3 n \tilde{\varepsilon}_n^2 \}) < g(u_n) e^{c_3 n \tilde{\varepsilon}_n^2} \right) , \\ &\leq \mathbb{P}_{\pi_*} \left(\mu \left(S_*(\tilde{\varepsilon}_n) \cap \{ \pi : \ell_n(\pi, X_{1:n}) - \ell_n(\pi_*, X_{1:n}) \geq -c_3 n \tilde{\varepsilon}_n^2 \}^c \right) > \mu(S_*(\tilde{\varepsilon}_n)) - g(u_n) e^{c_3 n \tilde{\varepsilon}_n^2} \right) . \end{aligned}$$

By Markov inequality,

$$\begin{aligned}
 \mathbb{P}_{\pi_*}(D_n < g(u_n)) &\leq \frac{\mathbb{E}_{\pi_*} [\mu(S_*(\tilde{\varepsilon}_n) \cap \{\pi : \ell_n(\pi, X_{1:n}) - \ell_n(\pi_*, X_{1:n}) \geq -c_3 n \tilde{\varepsilon}_n^2\}^c)]}{\mu(S_*(\tilde{\varepsilon}_n)) - g(u_n) e^{c_3 n \tilde{\varepsilon}_n^2}}, \\
 &\leq \frac{\int_{S_*(\tilde{\varepsilon}_n)} \mathbb{P}_{\pi_*}(\ell_n(\pi, X_{1:n}) - \ell_n(\pi_*, X_{1:n}) < -c_3 n \tilde{\varepsilon}_n^2) \mu(d\pi)}{\mu(S_*(\tilde{\varepsilon}_n)) - g(u_n) e^{c_3 n \tilde{\varepsilon}_n^2}}, \\
 &\leq \frac{\mu(S_*(\tilde{\varepsilon}_n))}{\mu(S_*(\tilde{\varepsilon}_n)) - g(u_n) e^{c_3 n \tilde{\varepsilon}_n^2}} \frac{c_\nu \log^2(\tilde{\varepsilon}_n)}{(c_3 - c_\nu)^2 n \tilde{\varepsilon}_n^2},
 \end{aligned}$$

where the last inequality follows from Proposition 1 and [15, Lemma 10]. Choosing $g : u \mapsto e^{-\tau u^2}/2$ with $\tau = (c_3 + c_1)/(\bar{c}/(2c) - 1)^2$ yields $g(u_n) e^{c_3 n \tilde{\varepsilon}_n^2} = e^{-c_1 n \tilde{\varepsilon}_n^2}/2$ and by H2-(c_0, c_1, c_2),

$$\mathbb{P}_{\pi_*}(D_n < g(u_n)) \leq \frac{2c_\nu \log^2(\tilde{\varepsilon}_n)}{(c_3 - c_\nu)^2 n \tilde{\varepsilon}_n^2}.$$

Then, with $c_3 = c_\nu + (c_2 - c_\nu - c_1 - c_\alpha)/2 > c_\nu$,

$$\frac{\mu(\Pi_n^c)}{\alpha_n g(u_n)} \leq \frac{2}{\alpha_n} e^{[(c_3 + c_1 - c_2)] n \tilde{\varepsilon}_n^2} = o(1),$$

which concludes the proof.

B Forgetting properties of the loglikelihood

Lemma 6. Assume that H1 holds. For any $i \geq 1$, conditionally on $X_{1:i}$, $(V_k)_{k \geq 1}$ is a Markov chain. Its transition kernels $(K_{\pi, k, i}^{V|X})_{k \geq 1}$ are such that, for all $1 \leq k \leq i$, there exists a measure $\mu_{k, i}$ satisfying for all measurable set A :

$$K_{\pi, k, i}^{V|X}(V_k, A) = \mathbb{P}_\pi(V_{k+1} \in A | V_{1:k}, X_{1:i}) = \mathbb{P}_\pi(V_{k+1} \in A | V_k, X_{k:i}) \geq \nu \mu_{k, i}(A),$$

where $\mu_{i, i} = \pi$ and for $1 \leq k < i$,

$$\mu_{k, i}(A) = \frac{\int \mathbb{1}_A(v_{k+1}) \pi(dv_{k+1}) \mathbb{P}_\pi(X_{k+1:i} | v_{k+1})}{\int \pi(dv_{k+1}) \mathbb{P}_\pi(X_{k+1:i} | v_{k+1})}.$$

On the other hand, for all $k \geq i + 1$ and all measurable set A ,

$$K_{\pi, k, i}^{V|X}(V_k, A) = \mathbb{P}_\pi(V_{k+1} \in A | V_{1:k}, X_{1:i}) = \pi(A).$$

Proof. The proof is a classical result for strong mixing latent data models, it follows closely [11]. See also [7] for a proof in the framework described by Figure 2. \square

Lemma 7. Assume that H1 holds. For all distribution π and all $i \geq 1$, $\ell \geq 0$,

$$|\log \mathbb{P}_\pi(x_i | x_{1:i-1}) - \log \mathbb{P}_\pi(x_i | x_{-\ell:i-1})| \leq \nu^{-1} (1 - \nu)^{i-1}.$$

There exists a function ℓ_π such that, for all distribution π' , $\mathbb{P}_{\pi'}$ -a.s. and in $L^1(\mathbb{P}_{\pi'})$,

$$\frac{1}{n} \ell_n(\pi, X_{1:n}) \xrightarrow{n \rightarrow \infty} \mathbb{L}_{\pi'}(\pi) = \mathbb{E}_{\pi'}[\ell_\pi(X)]$$

and

$$\left| \mathbb{E}_{\pi'} \left[\frac{1}{n} \ell_n(\pi, X_{1:n}) \right] - \mathbb{L}_{\pi'}(\pi) \right| \leq \frac{1}{n\nu^2}.$$

Proof. Let $x \in \mathcal{X}^{\mathbb{Z}}$. For all $i \geq 2$ and all $\ell \geq 0$,

$$\begin{aligned} \mathbb{P}_\pi(x_i | x_{1:i-1}) &= \int \mathbb{P}_\pi(dv_1 | x_{1:i-1}) \left(\prod_{k=1}^{i-1} K_{\pi, k, i-1}^{V|X}(v_k, dv_{k+1}) \right) \pi(dv_{i+1}) K(x_i, v_i, v_{i+1}), \\ \mathbb{P}_\pi(x_i | x_{-\ell:i-1}) &= \int \mathbb{P}_\pi(dv_1 | x_{-\ell:i-1}) \left(\prod_{k=1}^{i-1} K_{\pi, k, i-1}^{V|X}(v_k, dv_{k+1}) \right) \pi(dv_{i+1}) K(x_i, v_i, v_{i+1}), \end{aligned}$$

as $V_{2:i+1}$ is independent of $X_{-\ell:0}$ conditionally on V_1 . Therefore, by Lemma 6,

$$|\mathbb{P}_\pi(x_i | x_{1:i-1}) - \mathbb{P}_\pi(x_i | x_{-\ell:i-1})| \leq (1 - \nu)^{i-1}$$

and, since $\mathbb{P}_\pi(x_i | x_{1:i-1}) \wedge \mathbb{P}_\pi(x_i | x_{-\ell:i-1}) \geq \nu$,

$$|\log \mathbb{P}_\pi(x_i | x_{1:i-1}) - \log \mathbb{P}_\pi(x_i | x_{-\ell:i-1})| \leq \nu^{-1} (1 - \nu)^{i-1}.$$

Similarly, for $i = 1$,

$$|\log \mathbb{P}_\pi(x_1) - \log \mathbb{P}_\pi(x_1 | x_{-\ell:0})| \leq \nu^{-1}.$$

Let $(X_i, V_{i+1})_{i \in \mathbb{Z}}$ be a process with distribution $\mathbb{P}_{\pi'}$. For all $i \in \mathbb{Z}$,

$$|\log \mathbb{P}_\pi(X_i | X_{1:i-1}) - \log \mathbb{P}_\pi(X_i | X_{-\ell:i-1})| \leq \nu^{-1} (1 - \nu)^{i-1}.$$

Define the shift operator ϑ on $\mathcal{X}^{\mathbb{Z}}$ by $(\vartheta x)_i = x_{i+1}$ for all $i \in \mathbb{Z}$ and all $x \in \mathcal{X}^{\mathbb{Z}}$. There exists a function ℓ_π such that for all $i \in \mathbb{Z}$ the sequence $(\log \mathbb{P}_\pi(X_i | X_{-\ell:i-1}))_{\ell \geq 0}$ converges $\mathbb{P}_{\pi'}$ -a.s. to $\ell_\pi(\vartheta^i X)$. Then, when ℓ grows to ∞ , this yields

$$|\log \mathbb{P}_\pi(X_i | X_{1:i-1}) - \ell_\pi(\vartheta^i X)| \leq \nu^{-1} (1 - \nu)^{i-1}.$$

Then,

$$\left| \frac{1}{n} \ell_n(\pi, X_{1:n}) - \frac{1}{n} \sum_{i=1}^n \ell_\pi(\vartheta^i X) \right| = \left| \frac{1}{n} \sum_{i=1}^n \{ \log \mathbb{P}_\pi(X_i | X_{1:i-1}) - \ell_\pi(\vartheta^i X) \} \right| \leq \frac{1}{n\nu^2}$$

and the proof of the second claim is completed with the ergodic theorem [1, Theorem 24.1]. Remark that for all $i \geq 1$, $\mathbb{E}_{\pi'}[\ell_\pi(\vartheta^i X)] = \mathbb{L}_{\pi'}(\pi)$, the loglikelihood of $X_{1:n}$ also satisfies

$$\left| \mathbb{E}_{\pi'} \left[\frac{1}{n} \ell_n(\pi, X_{1:n}) \right] - \mathbb{L}_{\pi'}(\pi) \right| = \left| \mathbb{E}_{\pi'} \left[\frac{1}{n} \sum_{i=1}^n \{ \log \mathbb{P}_\pi(x_i | x_{1:i-1}) - \ell_\pi(\vartheta^i X) \} \right] \right| \leq \frac{1}{n\nu^2}.$$

□

Lemma 8. Assume that *H1* holds. For all $x_{1:n} \in \mathcal{X}^n$, all distributions $\pi, \pi' \in \Pi$ and any $1 \leq i \leq n$,

$$|\mathbb{P}_\pi(x_i|x_{1:i-1}) - \mathbb{P}_{\pi'}(x_i|x_{1:i-1})| \leq 2\nu^{-2}(2 + \nu^{-1})\|\pi - \pi'\|_{\text{tv}}.$$

Proof. When $i = 1$,

$$\mathbb{P}_\pi(x_1) - \mathbb{P}_{\pi'}(x_1) = \int \{\pi \otimes \pi(dv_{1:2}) - \pi' \otimes \pi'(dv_{1:2})\} \mathbb{K}(x_1, v_1, v_2).$$

Thus $|\mathbb{P}_\pi(x_1) - \mathbb{P}_{\pi'}(x_1)| \leq 2\|\pi - \pi'\|_{\text{tv}}$. For all $2 \leq i \leq n$,

$$\mathbb{P}_\pi(x_i|x_{1:i-1}) - \mathbb{P}_{\pi'}(x_i|x_{1:i-1}) = \sum_{\ell=1}^{i+1} \{\mathbb{P}_\ell(x_i|x_{1:i-1}) - \mathbb{P}_{\ell+1}(x_i|x_{1:i-1})\},$$

where \mathbb{P}_ℓ is the joint distribution of $(X_{1:i}, V_{1:i+1})$ when $(V_1, \dots, V_{\ell-1})$ are i.i.d. with distribution π' and (V_ℓ, \dots, V_{i+1}) are i.i.d. with distribution π . Then, for all $1 \leq \ell \leq i-1$,

$$\begin{aligned} \mathbb{P}_\ell(x_i|x_{1:i-1}) &= \int \mathbb{P}_\ell(dv_{\ell+1}|x_{1:i-1}) \left(\prod_{k=\ell+1}^i K_{\pi,k,i-1}^{V|X}(v_k, dv_{k+1}) \right) \mathbb{K}(x_i, v_i, v_{i+1}), \\ &= \int \mathbb{P}_\ell(dv_{\ell+1}|x_{1:i-1}) \left(\prod_{k=\ell+1}^{i-1} K_{\pi,k,i-1}^{V|X}(v_k, dv_{k+1}) \right) \pi(dv_{i+1}) \mathbb{K}(x_i, v_i, v_{i+1}). \end{aligned}$$

Therefore, for all $1 \leq \ell \leq i-1$, by Lemma 6,

$$|\mathbb{P}_\ell(x_i|x_{1:i-1}) - \mathbb{P}_{\ell+1}(x_i|x_{1:i-1})| \leq (1 - \nu)^{i-\ell-1} \|\mathbb{P}_\ell(\cdot|x_{1:i-1}) - \mathbb{P}_{\ell+1}(\cdot|x_{1:i-1})\|_{\text{tv}},$$

where $\mathbb{P}_\ell(\cdot|x_{1:i-1})$ is the distribution of $V_{\ell+1}$ conditionally on $\{X_{1:i-1} = x_{1:i-1}\}$ when $(V_1, \dots, V_{\ell-1})$ are i.i.d. with distribution π' and (V_ℓ, \dots, V_{i+1}) are i.i.d. with distribution π . We first show that

$$\|\mathbb{P}_\ell(\cdot|x_{1:i-1}) - \mathbb{P}_{\ell+1}(\cdot|x_{1:i-1})\|_{\text{tv}} \leq 2\nu^{-2}\|\pi - \pi'\|_{\text{tv}}.$$

Write, for all $1 \leq \ell \leq i-1$,

$$L_\ell^{x_{1:i-1}}(dv_{1:i}) = \prod_{m=1}^{\ell-1} \pi'(dv_m) \prod_{m=\ell}^i \pi(dv_m) \prod_{m=1}^{i-1} \mathbb{K}(x_m, v_m, v_{m+1}). \quad (35)$$

Then, for any function f on \mathcal{V} ,

$$\int f(v_{\ell+1}) \mathbb{P}_\ell(dv_{\ell+1}|x_{1:i-1}) = \frac{\int f(v_{\ell+1}) L_\ell^{x_{1:i-1}}(dv_{1:i})}{\int L_\ell^{x_{1:i-1}}(dv_{1:i})}.$$

Therefore,

$$\begin{aligned} &\int f(v_{\ell+1}) \{\mathbb{P}_\ell(dv_{\ell+1}|x_{1:i-1}) - \mathbb{P}_{\ell+1}(dv_{\ell+1}|x_{1:i-1})\} \\ &= \int f(v_{\ell+1}) \left(\frac{L_\ell^{x_{1:i-1}}(dv_{1:i})}{\int L_\ell^{x_{1:i-1}}(dv_{1:i})} - \frac{L_{\ell+1}^{x_{1:i-1}}(dv_{1:i})}{\int L_{\ell+1}^{x_{1:i-1}}(dv_{1:i})} \right), \\ &= \int f(v_{\ell+1}) \frac{L_\ell^{x_{1:i-1}}(dv_{1:i}) - L_{\ell+1}^{x_{1:i-1}}(dv_{1:i})}{\int L_\ell^{x_{1:i-1}}(dv_{1:i})} \\ &\quad + \int f(v_{\ell+1}) \frac{L_{\ell+1}^{x_{1:i-1}}(dv_{1:i})}{\int L_{\ell+1}^{x_{1:i-1}}(dv_{1:i})} \frac{\int [L_{\ell+1}^{x_{1:i-1}}(dv_{1:i}) - L_\ell^{x_{1:i-1}}(dv_{1:i})]}{\int L_\ell^{x_{1:i-1}}(dv_{1:i})}. \end{aligned}$$

Thus, for any function f on \mathcal{V} such that $\|f\|_\infty < 1$,

$$\left| \int f(v_{\ell+1}) \{ \mathbb{P}_\ell(dv_{\ell+1}|x_{1:i-1}) - \mathbb{P}_{\ell+1}(dv_{\ell+1}|x_{1:i-1}) \} \right| \leq 2 \frac{|\int \{L_\ell^{x_{1:i-1}}(dv_{1:i}) - L_{\ell+1}^{x_{1:i-1}}(dv_{1:i})\}|}{\int L_\ell^{x_{1:i-1}}(dv_{1:i})}. \quad (36)$$

By (35), $1 \leq \ell \leq i-1$,

$$\begin{aligned} & \left| \int \{L_\ell^{x_{1:i-1}}(dv_{1:i}) - L_{\ell+1}^{x_{1:i-1}}(dv_{1:i})\} \right| \\ &= \left| \int \prod_{m=1}^{\ell-1} \pi'(dv_m) \{ \pi(dv_\ell) - \pi'(dv_\ell) \} \prod_{m=\ell+1}^i \pi(dv_m) \prod_{m=1}^{i-1} \mathsf{K}(x_m, v_m, v_{m+1}) \right|. \end{aligned}$$

As K is upper bounded by 1,

$$\begin{aligned} \left| \int \{L_\ell^{x_{1:i-1}}(dv_{1:i}) - L_{\ell+1}^{x_{1:i-1}}(dv_{1:i})\} \right| &\leq \left(\int \prod_{m=1}^{\ell-1} \pi'(dv_m) \prod_{m=1}^{\ell-2} \mathsf{K}(x_m, v_m, v_{m+1}) \right) \\ &\quad \times \|\pi - \pi'\|_{\text{tv}} \left(\int \prod_{m=\ell+1}^i \pi(dv_m) \prod_{m=\ell+1}^{i-1} \mathsf{K}(x_m, v_m, v_{m+1}) \right). \end{aligned}$$

Similarly, since K is respectively lower bounded by ν ,

$$\begin{aligned} \int L_\ell^{x_{1:i-1}}(dv_{1:i}) &\geq \left(\int \prod_{m=1}^{\ell-1} \pi'(dv_m) \prod_{m=1}^{\ell-2} \mathsf{K}(x_m, v_m, v_{m+1}) \right) \\ &\quad \times \nu^2 \left(\int \prod_{m=\ell+1}^i \pi(dv_m) \prod_{m=\ell+1}^{i-1} \mathsf{K}(x_m, v_m, v_{m+1}) \right). \end{aligned}$$

Plugging these bounds in (36) yields, for $1 \leq \ell \leq i-1$,

$$\left| \int f(v_{\ell+1}) \{ \mathbb{P}_\ell(dv_{\ell+1}|x_{1:i-1}) - \mathbb{P}_{\ell+1}(dv_{\ell+1}|x_{1:i-1}) \} \right| \leq 2\nu^{-2} \|\pi - \pi'\|_{\text{tv}}.$$

Then, by bounding similarly the two last terms of the telescoping sum,

$$\mathbb{P}_\pi(x_i|x_{1:i-1}) - \mathbb{P}_{\pi'}(x_i|x_{1:i-1}) = 2\nu^{-2} \|\pi - \pi'\|_{\text{tv}} \left\{ \sum_{\ell=1}^{i-1} (1-\nu)^{i-\ell-1} + 2 \right\} \leq 2(2+\nu^{-1})\nu^{-2} \|\pi - \pi'\|_{\text{tv}}.$$

□

C Forgetting properties of the predictive filter

For all $k \geq 1$, η_k^π is the predictive filter at time k when the law of the hidden states is π : $\eta_1^\pi = \pi$ and for $k \geq 2$,

$$\eta_k^\pi : A \mapsto \mathbb{P}_\pi(V_k \in A | X_{1:k-1}).$$

By applying (25) recursively, for all $p \geq 1$, define the function Φ_p by:

$$\eta_{p+1}^\pi = \Phi(X_p, \eta_p^\pi; \pi) = \Phi(X_p, \Phi(X_{p-1}, \eta_{p-1}^\pi; \pi); \pi) = \dots = \Phi_p(X_{1:p}, \eta_1^\pi; \pi),$$

with the convention $\Phi_1 = \Phi$. Lemma 9 establishes the exponential forgetting of the prediction filter. Its proof follows closely [9, Proposition 1] in the case of general hidden Markov models.

Lemma 9. *Assume that H1 holds. For all $p \geq 1$, all $x_{1:p} \in \mathcal{X}^p$ and all probability distributions η, η', π on \mathcal{V} ,*

$$\|\Phi_p(x_{1:p}, \eta; \pi) - \Phi_p(x_{1:p}, \eta'; \pi)\|_{\text{tv}} \leq (1 - \nu)^p.$$

Proof. By definition, $\Phi_p(x_{1:p}, \eta; \pi)$ is the predictive distribution of V_{p+1} given $\{X_{1:p} = x_{1:p}\}$ when $V_1 \sim \eta$ and $V_k \sim \pi$ for $2 \leq k \leq p+1$. By Lemma 6,

$$\int \Phi_p(x_{1:p}, \eta; \pi)(dv_{p+1})h(v_{p+1}) = \int \phi_{\eta, \pi, p}(dv_1) \prod_{k=1}^p K_{\pi, k, p}^{V|X}(v_k, dv_{k+1})h(v_{p+1}),$$

where $\phi_{\eta, \pi, p}$ is the conditional distribution of V_1 given $\{X_{1:p} = x_{1:p}\}$ when $V_1 \sim \eta$ and $V_k \sim \pi$ for $2 \leq k \leq p+1$. Therefore, by Lemma 6,

$$\|\Phi_p(x_{1:p}, \eta; \pi) - \Phi_p(x_{1:p}, \eta'; \pi)\|_{\text{tv}} \leq (1 - \nu)^p \|\phi_{\eta, \pi, p} - \phi_{\eta', \pi, p}\|_{\text{tv}},$$

which concludes the proof. \square

Lemma 10. *Assume H1 holds. Then, for all $r = (v, x, \eta, \eta_\star) \in \mathcal{V} \times \mathcal{X} \times \mathcal{S}^+ \times \mathcal{S}^+$, $r' = (v', x', \eta', \eta'_\star) \in \mathcal{V} \times \mathcal{X} \times \mathcal{S}^+ \times \mathcal{S}^+$ and all $p \geq 1$,*

$$|Q_\pi^p \mathbf{h}(r) - Q_\pi^p \mathbf{h}(r')| \leq 6\nu^{-1}(1 - \nu)^{p-2},$$

where \mathbf{h} and Q_π are defined in (24) and (26).

Proof. For all measurable function h , $r = (v, x, \eta, \eta_\star) \in \mathcal{X} \times \mathcal{V} \times \mathcal{S}^+ \times \mathcal{S}^+$, $r' = (v', x', \eta', \eta'_\star) \in \mathcal{X} \times \mathcal{V} \times \mathcal{S}^+ \times \mathcal{S}^+$ and all $p \geq 1$,

$$|Q_\pi^p h(r) - Q_\pi^p h(r')| \leq |Q_\pi^p h(v, x, \eta, \eta_\star) - Q_\pi^p h(v, x, \eta', \eta'_\star)| + |Q_\pi^p h(v, x, \eta', \eta'_\star) - Q_\pi^p h(v', x', \eta', \eta'_\star)|. \quad (37)$$

With the convention $v_2 = v$, for all $p \geq 1$,

$$Q_\pi^p h(r) = \sum_{x_{2:p+1} \in \mathcal{X}^p} \int h(v_{p+2}, x_{p+1}, \Phi_p((x, x_{2:p}), \eta; \pi), \Phi_p((x, x_{2:p}), \eta_\star; \pi_\star)) \times \prod_{k=3}^{p+2} K(x_{k-1}, v_{k-1}, v_k) \pi^{\otimes p}(dv_{3:p+2}).$$

By H1, the function \mathbf{h} defined in (24) satisfies:

$$|\mathbf{h}(v, x, \eta, \eta_\star) - \mathbf{h}(v, x, \eta', \eta'_\star)| \leq \nu^{-1} (\|\eta - \eta'\|_{\text{tv}} + \|\eta_\star - \eta'_\star\|_{\text{tv}}). \quad (38)$$

The exponential forgetting property of the predictive filter given by Lemma 9 yields

$$\|\Phi_{p-1}(x_{2:p}, \Phi(x, \eta; \pi); \pi) - \Phi_{p-1}(x_{2:p}, \Phi(x, \eta'; \pi); \pi)\|_{\text{tv}} \leq (1 - \nu)^{p-1} \quad (39)$$

and, as $\Phi_p((x, x_{2:p}), \eta; \pi) = \Phi_{p-1}(x_{2:p}, \Phi(x, \eta; \pi); \pi)$,

$$|Q_\pi^p h(v, x, \eta, \eta_\star) - Q_\pi^p h(v, x, \eta', \eta'_\star)| \leq \frac{2}{\nu} (1 - \nu)^{p-1}.$$

For the second term of (37), write for $p \geq 2$,

$$\Phi_p((x, x_{2:p}), \eta; \pi) = \Phi_{p-2}(x_{3:p}, \Phi_2((x, x_2), \eta; \pi); \pi)$$

and for any ϕ and ϕ_\star in \mathcal{P} ,

$$\begin{aligned} h(v_{p+2}, x_{p+1}, \Phi_p((x, x_{2:p}), \eta; \pi), \Phi_p((x, x_{2:p}), \eta_\star; \pi_\star)) = \\ h(v_{p+2}, x_{p+1}, \Phi_{p-2}(x_{3:p}, \Phi_2((x, x_2), \eta; \pi); \pi), \Phi_{p-2}(x_{3:p}, \Phi_2((x, x_2), \eta_\star; \pi_\star); \pi_\star)) \\ - h(v_{p+2}, x_{p+1}, \Phi_{p-2}(x_{3:p}, \phi; \pi), \Phi_{p-2}(x_{3:p}, \phi_\star; \pi_\star)) \\ + h(v_{p+2}, x_{p+1}, \Phi_{p-2}(x_{3:p}, \phi; \pi), \Phi_{p-2}(x_{3:p}, \phi_\star; \pi_\star)). \end{aligned}$$

Then using that

$$\begin{aligned} \sum_{x_{2:p+1} \in \mathcal{X}^p} \int h(v_{p+2}, x_{p+1}, \Phi_{p-2}(x_{3:p}, \phi; \pi), \Phi_{p-2}(x_{3:p}, \phi_\star; \pi_\star)) \prod_{k=4}^{p+2} K(x_{k-1}, v_{k-1}, v_k) K(x_2, v, v_3) \pi^{\otimes p}(\mathrm{d}v_{3:p+2}) \\ = \sum_{x_{2:p+1} \in \mathcal{X}^p} \int h(v_{p+2}, x_{p+1}, \Phi_{p-2}(x_{3:p}, \phi; \pi), \Phi_{p-2}(x_{3:p}, \phi_\star; \pi_\star)) \prod_{k=4}^{p+2} K(x_{k-1}, v_{k-1}, v_k) K(x_2, v', v_3) \pi^{\otimes p}(\mathrm{d}v_{3:p+2}) \end{aligned}$$

and (38) and (39),

$$\begin{aligned} |Q_\pi^p h(v, x, \eta', \eta'_\star) - Q_\pi^p h(v', x', \eta', \eta'_\star)| \\ \leq 2\nu^{-1} (1 - \nu)^{p-2} \sum_{x_{2:p+1} \in \mathcal{X}^p} \int (K(x_2, v, v_3) + K(x_2, v', v_3)) \prod_{k=4}^{p+2} K(x_{k-1}, v_{k-1}, v_k) \pi^{\otimes p}(\mathrm{d}v_{3:p+2}) \\ \leq 4\nu^{-1} (1 - \nu)^{p-2}. \end{aligned}$$

In the case $p = 1$,

$$|Q_\pi^p h(v, x, \eta', \eta'_\star) - Q_\pi^p h(v', x', \eta', \eta'_\star)| \leq 4\nu^{-1}.$$

□

D Concentration inequality for Markov chains with uniformly lower bounded transition kernel

Theorem 11 and its proof can be found in [10]. As this book is not published yet, the proof of this result is reproduced here for the sake of completeness. Let $\nu > 0$ and $(X, B(X))$ be a measurable

space. In this section, P denotes a Markov kernel defined on $(X, B(X))$ bounded from below by $\nu > 0$: there exists a probability measure η on $(X, B(X))$ such that for all $A \in B(X)$ and all $x \in X$,

$$P(x, A) \geq \nu \eta(A). \quad (40)$$

For any $x \in X$ (resp. any probability distribution π defined on $(X, B(X))$), let P_x (resp. P_π) be the distribution of a Markov chain $(X_n)_{n \geq 0}$ with Markov kernel P and such that $X_0 = x$ (resp. $X_0 \sim \pi$). The expectation w.r.t. P_x (resp. P_π) is denoted by E_x (resp. E_π). Let $\gamma \in \mathbb{R}_+^n$ and $B(\gamma)$ be the set of functions $f : X \rightarrow \mathbb{R}$ such that, for all $x, y \in X^n$, $k \in \{1, \dots, n\}$,

$$|f(x) - f(y)| \leq \sum_{k=1}^n \gamma_k \mathbb{1}_{x_k \neq y_k}.$$

Following for instance [3, Chapter 4], the Dobrushin coefficient of P can be defined as

$$\Delta(P) = \sup_{(x, x') \in X^2} \|P(x, \cdot) - P(x', \cdot)\|_{\text{tv}} = \sup_{\pi \neq \pi'} \frac{\|\pi P - \pi' P\|_{\text{tv}}}{\|\pi - \pi'\|_{\text{tv}}},$$

where for any distribution π on $(X, B(X))$ and all $A \in B(X)$, $\pi P(A) = \int \pi(dx) P(x, A)$. For all $n \geq 1$, define

$$D_n = \sum_{\ell=1}^n \left(\gamma_\ell + 2 \sum_{m=\ell+1}^n \gamma_m \Delta(P^{m-\ell}) \right)^2. \quad (41)$$

Theorem 11. *For any distribution π on $(X, B(X))$, any $\gamma \in \mathbb{R}_+^n$, any $f \in B(\gamma)$ and any $t > 0$,*

$$P_\pi(f(X_1, \dots, X_n) - E_\pi[f(X_1, \dots, X_n)] > t) \leq e^{-2t^2/D_n},$$

where D_n is defined by (41).

Proof. Assume without loss of generality that $E_\pi[f(X_1, \dots, X_n)] = 0$. For any $\ell \in \{1, \dots, n\}$, define

$$A_\ell = \gamma_\ell + 2 \sum_{m=\ell+1}^n \gamma_m \Delta(P^{m-\ell}).$$

For all $\ell \geq 0$, define $\mathcal{F}_\ell = \sigma(X_i, 1 \leq i \leq \ell)$ and $G_\ell = E_\pi[f(X_1, \dots, X_n) | \mathcal{F}_\ell]$ with the conventions $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $G_0 = 0$. Hence,

$$f(X_1, \dots, X_n) = \sum_{\ell=1}^n (G_\ell - G_{\ell-1}), \quad G_{\ell-1} = E[G_\ell | \mathcal{F}_{\ell-1}].$$

The pivotal step of the proof is to establish that for all $1 \leq \ell \leq n$ there exist $\mathcal{F}_{\ell-1}$ -measurable random variables a_ℓ and b_ℓ such that

$$a_\ell \leq G_\ell \leq b_\ell \quad \text{and} \quad b_\ell - a_\ell \leq A_\ell. \quad (42)$$

For all $1 \leq \ell \leq n$, $\ell \leq s \leq n-1$, $x_{1:\ell} \in X^\ell$ and $x_{s+1:n} \in X^{n-s}$ define

$$K_{\ell,s}(x_1, \dots, x_\ell, x_{s+1}, \dots, x_n) = \inf_{(u_{\ell+1}, \dots, u_s) \in X^{s-\ell}} f(x_1, \dots, x_\ell, u_{\ell+1}, \dots, u_s, x_{s+1}, \dots, x_n)$$

and

$$W_{\ell,s}(x_1, \dots, x_\ell, x_{s+1}, \dots, x_n) = K_{\ell,s}(x_1, \dots, x_\ell, x_{s+1}, \dots, x_n) - K_{\ell,s+1}(x_1, \dots, x_\ell, x_{s+2}, \dots, x_n).$$

Hence, $0 \leq W_{\ell,s} \leq \gamma_{s+1}$ and

$$G_\ell = K_{\ell,n}(X_1, \dots, X_\ell) + \sum_{s=\ell}^{n-1} \mathbb{E}_\pi[W_{\ell,s}(X_1, \dots, X_\ell, X_{s+1}, \dots, X_n) | \mathcal{F}_\ell].$$

Let X_1^*, \dots, X_n^* denote an independent copy of X_1, \dots, X_n so that

$$G_\ell = K_{\ell,n}(X_1, \dots, X_\ell) + \sum_{s=\ell}^{n-1} \mathbb{E}_\pi[W_{\ell,s}(X_1, \dots, X_\ell, X_{s+1}^*, \dots, X_n^*) | \mathcal{F}_\ell].$$

Defining $H_\ell = \mathbb{E}_\pi[f(X_0, \dots, X_\ell, X_{\ell+1}^*, \dots, X_n^*) | \mathcal{F}_\ell]$ yields

$$|G_\ell - H_\ell| \leq \sum_{s=\ell}^{n-1} \left| \mathbb{E}_\pi[W_{\ell,s}(X_1, \dots, X_\ell, X_{s+1}, \dots, X_n) | \mathcal{F}_\ell] - \mathbb{E}_\pi[W_{\ell,s}(X_1, \dots, X_\ell, X_{s+1}^*, \dots, X_n^*) | \mathcal{F}_\ell] \right|.$$

Write for all $y \in \mathbf{X}$,

$$h(y) = \mathbb{E}_\pi[W_{\ell,s}(X_1, \dots, X_\ell, y) | \mathcal{F}_\ell] \quad \text{and} \quad h'(y) = \mathbb{E}_y[h(X_1, \dots, X_{n-s})].$$

Then,

$$\begin{aligned} & \left| \mathbb{E}_\pi[W_{\ell,s}(X_1, \dots, X_\ell, X_{s+1}, \dots, X_n) | \mathcal{F}_\ell] - \mathbb{E}_\pi[W_{\ell,s}(X_1, \dots, X_\ell, X_{s+1}^*, \dots, X_n^*) | \mathcal{F}_\ell] \right| \\ &= \left| \mathbb{E}_{X_\ell}[h(X_{s+1-\ell}, \dots, X_{n-\ell})] - \mathbb{E}_{\pi^{P^\ell}}[h(X_{s+1-\ell}, \dots, X_{n-\ell})] \right|, \\ &= \int \pi^{P^\ell}(dy) (P^{s+1-\ell} h'(X_\ell) - P^{s+1-\ell} h'(y)), \\ &\leq \|h'\|_\infty \Delta(P^{s+1-\ell}). \end{aligned}$$

As $\|h'\|_\infty \leq \|W_{\ell,s}\|_\infty \leq \gamma_{s+1}$, we have

$$|G_\ell - H_\ell| \leq \sum_{s=\ell+1}^n \gamma_s \Delta(P^{s-\ell}).$$

Define, for all $1 \leq \ell \leq n$, the following $\mathcal{F}_{\ell-1}$ -measurable random variables

$$\begin{aligned} a'_\ell &= \inf_{x_\ell \in \mathbf{X}} \mathbb{E}[f(X_1, \dots, X_{\ell-1}, x_\ell, X_{\ell+1}^*, \dots, X_n^*) | \mathcal{F}_\ell], \\ b'_\ell &= \sup_{x_\ell \in \mathbf{X}} \mathbb{E}[f(X_1, \dots, X_{\ell-1}, x_\ell, X_{\ell+1}^*, \dots, X_n^*) | \mathcal{F}_\ell]. \end{aligned}$$

Since $a'_\ell \leq H_\ell \leq b'_\ell$ and $b'_\ell - a'_\ell \leq \gamma_\ell$, it follows that

$$a_\ell = a'_\ell - \sum_{s=\ell+1}^n \gamma_s \Delta(P^{s-\ell}) \quad \text{and} \quad b_\ell = b'_\ell + \sum_{s=\ell+1}^n \gamma_s \Delta(P^{s-\ell})$$

satisfy (42). By McDiarmid's inequality, for all $\lambda > 0$,

$$8 \log \mathbb{E}_\pi [e^{\lambda(G_\ell - G_{\ell-1})} | \mathcal{F}_{\ell-1}] \leq \lambda^2 A_\ell^2.$$

Then, by Markov's inequality, for any $s > 0$,

$$\mathbb{P}_\pi (f(X_1, \dots, X_n) > t) \leq e^{-\lambda t + \frac{\lambda^2}{8} \sum_{\ell=1}^n A_\ell^2} = e^{-\lambda t + \frac{\lambda^2 D_n}{8}}.$$

Applying this inequality with $\lambda = 4t/D_n$ concludes the proof. \square

Corollary 12. *For any distribution π on \mathbf{X} , any $\gamma \in \mathbb{R}_+^n$, any $f \in B(\gamma)$ and any $t > 0$,*

$$\mathbb{P}_\pi (f(X_1, \dots, X_n) - \mathbb{E}_\pi[f(X_1, \dots, X_n)] > t) \leq e^{-\frac{\nu^2}{5} \frac{t^2}{\sum_{\ell=1}^n \gamma_\ell^2}}.$$

Proof. For all $x \in \mathbf{X}$, define $Q(x, \cdot) = (1 - \nu)^{-1}(P(x, \cdot) - \nu\eta(\cdot))$ where ν and η are given by (40). For any $x \in \mathbf{X}$, $Q(x, \cdot)$ is a probability distribution and

$$P(x, \cdot) = (1 - \nu)Q(x, \cdot) + \nu\eta(\cdot).$$

Therefore, for all $(x, x') \in \mathbf{X}^2$,

$$\|P(x, \cdot) - P(x', \cdot)\|_{\text{tv}} = (1 - \nu)\|Q(x, \cdot) - Q(x', \cdot)\|_{\text{tv}} \leq 1 - \nu,$$

which yields, for all $q \geq 1$, $\Delta(P^q) \leq (\Delta(P))^q \leq (1 - \nu)^q$. By Theorem 11, for any distribution π on \mathbf{X} , any $\gamma \in \mathbb{R}_+^n$, any $f \in B(\gamma)$ and any $t > 0$,

$$\mathbb{P}_\pi (f(X_1, \dots, X_n) - \mathbb{E}_\pi[f(X_1, \dots, X_n)] > t) \leq e^{-\frac{2t^2}{\sum_{\ell=1}^n (\gamma_\ell + 2 \sum_{m=\ell+1}^n \gamma_m (1 - \nu)^{m-\ell})^2}}.$$

Note that by Cauchy-Schwarz inequality,

$$\begin{aligned} \sum_{\ell=1}^n \left(\gamma_\ell + 2 \sum_{m=\ell+1}^n \gamma_m (1 - \nu)^{m-\ell} \right)^2 &\leq 2 \sum_{\ell=1}^n \gamma_\ell^2 + 8 \sum_{\ell=1}^n \left(\sum_{m=\ell+1}^n \gamma_m (1 - \nu)^{m-\ell} \right)^2, \\ &\leq 2 \sum_{\ell=1}^n \gamma_\ell^2 + 8 \frac{1 - \nu}{\nu} \sum_{\ell=1}^n \left(\sum_{m=\ell+1}^n \gamma_m^2 (1 - \nu)^{m-\ell} \right), \\ &\leq 2 \left(1 + 4 \frac{(1 - \nu)^2}{\nu^2} \right) \sum_{\ell=1}^n \gamma_\ell^2, \end{aligned}$$

which concludes the proof. \square

References

- [1] P. Billingsley. *Probability and Measure*. Wiley, 1995.

- [2] R. Bradley and M. Terry. Rank analysis of incomplete block designs: I. the method of pair comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- [3] O. Cappe, É. Moulines, and T. Ryden. *Inference in hidden Markov models*. Springer, 2005.
- [4] D. Caron and A. Doucet. Efficient Bayesian inference for generalized Bradley-Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.
- [5] R. Chetrite, R. Diel, and M. Lerasle. The number of potential winners in bradley-terry models in random environment. *The Annals of Applied Probability*, 27(3):1372–1394, 2017.
- [6] P. Del Moral, A. Doucet, and S. Singh. A Backward Particle Interpretation of Feynman-Kac Formulae. *ESAIM M2AN*, 44(5):947–975, 2010.
- [7] R. Diel, S. Le Corff, and M. Lerasle. Learning latent structure of large random graphs. *arXiv:1707.01365*, 2018.
- [8] R. Douc, A. Garivier, É. Moulines, and J. Olsson. Sequential Monte Carlo smoothing for general state space hidden Markov models. *The Annals of Applied Probability*, 21(6):2109–2145, 2011.
- [9] R. Douc and C. Matias. Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7(3):381–420, 2001.
- [10] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov Chains*. Springer, 2018.
- [11] R. Douc, É. Moulines, and T. Ryden. Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *The Annals of Statistics*, 32(5):2254–2304, 2004.
- [12] C. Dubarry and S. Le Corff. Fast computation of smoothed additive functionals in general state-space models. *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 197–200, 2011.
- [13] C. Dubarry and S. Le Corff. Nonasymptotic deviation inequalities for smoothed additive functionals in nonlinear state-space models with applications to parameter estimation. *Bernoulli*, 19(5B):2222–2249, 2013.
- [14] S. Ghosal. Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.*, 29(5):1264–1280, 2001.
- [15] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.*, 35(1):192–223, 2007.
- [16] S. Ghosal and A. van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2017.
- [17] S. J. Godsill, A. Doucet, and M. West. Monte Carlo smoothing for non-linear time series. *Journal of the American Statistical Association*, 50:438–449, 2004.

- [18] N. Gordon, D. Salmond, and A.F. Smith. Novel approach to nonlinear/non-Gaussian bayesian state estimation. *IEE Proc. F, Radar Signal Process*, 140:107–113, 1993.
- [19] D.R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.
- [20] H. Joe. Extended use of paired comparison models, with application to chess rankings. *Journal of the Royal Statistical Society: Series C*, 39(1):85–93, 1990.
- [21] G. Kitagawa. Monte-Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 1:1–25, 1996.
- [22] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d’Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].
- [23] W. Kruijer, J. Rousseau, and A. van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257, 2010.
- [24] J. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.
- [25] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.
- [26] M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.
- [27] J. Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *The Annals of Statistics*, 38:146–180, 2010.
- [28] J. Rousseau. On the frequentist properties of Bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3:211–231, 2016.
- [29] W. Shen, S. Tokdar, and S. Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.
- [30] G. Simons and Y.-C. Yao. Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, 1999.
- [31] C. Sire and S. Redner. Understanding baseball team standings and streaks. *Eur. Phys. J. B*, 67:473–481, 2009.
- [32] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.
- [33] V.N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998. A Wiley-Interscience Publication.

- [34] É. Vernet. Non parametric hidden markov models with finite state space: Posterior concentration rates. *arXiv:1511.08624*, 2015.
- [35] S.G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54, 2007.
- [36] M.J. Whiting, D.M. Stuart-Fox, O’Connor D., D. Firth, N.C. Bennett, and S.P. Bloomberg. Ultraviolet signals ultra-aggression in a lizard. *Animal behaviour*, 72:353–363, 2006.
- [37] T. Yan, Y. Yang, and J. Xu. Sparse paired comparisons in the Bradley-Terry model. *Statist. Sinica*, 22(3):1305–1318, 2012.
- [38] C. Yau, O. Papaspiliopoulos, G.O. Roberts, and C. Holmes. Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B*, 73:37–57, 2011.
- [39] E. Zermelo. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Math. Z.*, 29(1):436–460, 1929.